UNDERSTANDING PROTEIN STRUCTURE AND DYNAMICS: FROM COMPARATIVE MODELING POINT OF VIEW TO DYNAMICAL PERSPECTIVES

A Thesis
Presented to
The Academic Faculty

by

Gungor Ozer

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Chemistry and Biochemistry

Georgia Institute of Technology
May 2011

UNDERSTANDING PROTEIN STRUCTURE AND DYNAMICS: FROM COMPARATIVE MODELING POINT OF VIEW TO DYNAMICAL PERSPECTIVES

Approved by:

Dr. Rigoberto Hernandez, Advisor
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Jean-Luc Brédas
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Stephen Harvey
School of Biology
Georgia Institute of Technology

Dr. C. David Sherrill
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Joseph Perry
School of Chemistry and Biochemistry
Georgia Institute of Technology

Date Approved: 31 March 2011

To my parents,

Fidan and Feyzi Özer,

for their continual support and counsel.

ACKNOWLEDGEMENTS

This thesis would not have been possible without constant support and encouragement from several individuals who in more than one way have contributed their valuable assistance throughout the whole study.

First and foremost, I would like to extend my utmost gratitude to my advisor, Professor Rigoberto Hernandez, who has patiently supported me at all points along the path. His knowledge, experience, wisdom and guidance have been the steering force to my success. He has been a great mentor and I will always be grateful. I would like to thank my collaborators whom I enjoyed working with: Dr. Stephen Quirk who was involved at every step of my work; Dr. Edward F. Valeev who completed some of the early simulations; and finally Dr. Shi Zhong who addressed my numerous questions and concerns even when he was not working in the Hernandez group.

I would like to acknowledge the members of my thesis committee who have been very involved in following my academic and professional development and instructive to excel my skills: Professor Stephen Harvey for sharing his pearls of wisdom in many aspects including theory and future directions, and presenting and publishing my findings effectively, Professor Jean-Luc Brédas for his patient efforts to improve my work even at the last moment, Professor Joseph Perry for periodically evaluating my research progress face to face, and Professor David Sherrill for always being available for my inquiries. I would also like to express my deepest appreciation to all the staff and faculty at the School of Chemistry and Biochemistry especially to Professor Christine Payne, Dr. Leigh Bottomley and Dr. Cam Tyson for their invaluable mentorship throughout my term at Georgia Tech.

I would be remiss if I didn't mention my coworkers and officemates that I had

the pleasure to work with first in Boggs and then in MoSE buildings — particularly all former and current members of the Hernandez, Sherrill and Brédas groups. I feel privileged to be a part of the CCMST team at Georgia Tech. I would like to extend my especial gratitude to Dr. Alex Popov, Ashley Tucker, Matt Hagy, Dr. Eli Hershkovits, Galen Craven, Jay Foley, Dr. Jeremy Moix, Denise Enekwa and Megan Damm for the team spirit; to Dr. Ashley Ringer, Dr. Berhane Temelso, Edward Hohenstein, Dr. John Sears, Michael Marshall and Steve Arnstein for contributing to the joyful work environment.

Finally, I would like to thank all my friends and family for their continuous support and solidarity. Special recognition goes to Seyhan Salman, Onur Kececigil, Haldun Kececigil, Serdar Ozdemir, Mustafa Cenk, Menderes Iskin and other members of the Alaturka FC, Anthony Appleton, Peter Hotchkiss, Ozgul Persil, Mehmet Cetinkol, Gozde Guler, Asli Ovat, Mustafa Burak Boz, Altug Kasali, Alper Akanser, Matthew Trotter and other participants of our fantasy league—GTsmarties. Of course, none of this would have been possible without Rezarta Bilali who was always there for and with me at every up and down of this whole process. Lastly I would like to express my gratitude to my family for their counsel and encouragement throughout my PhD study. I truly appreciate the never-ending support from my father Feyzi, my mother Fidan and my lovely sisters Gunay and Pinar. They have sacrificed so much for me and I will be forever indebted.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this thesis, we have advanced a set of distinct bioinformatic and computational tools to address the structure and function of proteins. Using data mining of the protein data bank (PDB), we have collected statistics connecting the propensity between the protein sequence and the secondary structure. This new tool has enabled us to evaluate new structures as well as a family of structures. A comparison of the wild type staphylococcal nuclease to various mutants using the proposed tool has indicated long-range conformational deviations spatially distant from the mutation point. The energetics of protein unfolding has been studied in terms of the forces observed in molecular dynamics simulations. An adaptive integration of the steered molecular dynamics is proposed to reduce ground state dominance by the rare low energy trajectories on the estimated free energy profile. The proposed adaptive algorithm is utilized to reproduce the potential of mean force of the stretching of decaalanine in vacuum at lower computational cost It is then used to construct the potential of mean force of this transition in solvent for the first time and to observe the hydration effect on the helix-coil transformation. Adaptive steered molecular dynamics is also implemented to obtain the free energy change during the unfolding of neuropeptide Y and to confirm that the monomeric form of neuropeptide Y adopts halical-hairpin like pancreatic-polypeptide fold.

The structure propensity of predicted or experimentally determined protein structures as well as family of structures is examined via a comparative modeling approach. The evaluation tool developed within the framework of this thesis utilizes a novel complementary checking function, D_2Check, recently developed by our group. We have extended the D_2Check analysis from the protein scale to that of the amino acids so to

identify typical and atypical values of dihedral angles about a single residue ($\phi - \psi$) or those about two adjacent residues (ψ_i-ϕ_{i+1}). At the residue level, a compact graphical representation is introduced to project dihedral angle compatibility of every amino acid (residual D_2 score) of a given structure onto a color-coded strip. The color strip can be used to visually identify the typicality or atypicality of a given structure. This is possible since a particular structure is observed to be atypical only when most of its residues have atypical D_2 values (i.e. adopt unlikely dihedral angles). One can visually observe the likelihood of residue dihedral angles through a representation using color intensity to assess the propensities of the overall protein structure at a glance. The color strip difference strip, on the other hand, can be used to analyze structural similarities/differences among protein families and structural effects of mutations. The color strip difference analysis of wild-type Staphylococcal nuclease (STN) for various LYS116 mutants has provided visual identification of the mutation site as well as other key sites that had been claimed as STN's biologically active regions. The D_2Check methodology has been integrated into a web server (http://www.d2check.gatech.edu) to make the D2 code available to the scientific community. The server includes both protein level and residue level analysis and provides users with raw data as well as consequent graphs such as color strips, Ramachandran plots, position of the overall D2 score in the D2 distribution.

All-atom molecular dynamics (MD) has been extensively used to study motion of biomolecules (e.g. protein folding). However, conformational sampling of protein folding and unfolding events at the atomic scale requires substantial amount of computation and, this, is usually limited to shorter timescales compared to the real life events. Many methodologies, such as steered molecular dynamics (SMD), have been developed within the framework of molecular dynamics, to accelerate these events. SMD works by applying a series of time-dependent external forces on the system, for example on a model protein along a preselected unfolding pathway. When the

system is driven through a path via external forces, it moves away from equilibrium. Jarzynski's inequality relates the applied force (i.e. work) to the potential of mean force (i.e. equilibrium free energy). For small systems with low energetic barriers, SMD in combination with the Jarzynski's nonequilibirum work relation yields accurate estimate of the potential of mean force (PMF) using a computationally accessible number of trajectories. For larger systems with higher energy barriers driven along extended paths, on the other hand, the applied force (thus required work) fluctuates dramatically across a very large range. Only the lowest energy trajectories dominate the PMF, and convergence would be achieved only through the determination of a prohibitively large number of trajectories. This can be surmounted by (i) increasing the sample size (as many as millions of realizations), (ii) decreasing the pulling velocity (as low as reversible velocity), or (iii) equilibrating the system at short intervals. All of these plausible solutions will, however, increase the amount of computation dramatically. This thesis presents a staged integration of the SMD methodology — adaptive steered molecular dynamics— that can be used to obtain a converged PMF efficiently. Each stage (or step) is designed to be short enough so that the work distribution exhibits good statistics and thus the corresponding PMF represents most of the generated trajectories. Adaptive SMD has been used to investigate the helix-coil transition of decaalanine in vacuum and in solvent and unfolding of several neurotransmitters (i.e. neuropeptide Y—NPY, peptide YY—PYY— and several PYY mutants) in solvent. The PMF along the stretching of decaalanine in vacuum was reproduced using adaptive SMD at much lower computational cost compared to conventional SMD. In solvent stretching of decaalanine using adaptive SMD has yielded an overall lowering of the PMF due to the stabilizing effect of the neighboring water molecules. The hydration effect is also confirmed analyzing the intra-peptide and peptide-water hydrogen bond counts. Adaptive SMD has also been used to calculate the PMF along the unfolding pathway of neuropeptide Y (NPY). Using this PMF

and the activation energy barrier observed on it, the transition state rate of the unfolding of NPY has been calculated. The results show that monomeric NPY adopts the pancreatic-polypeptide fold as proposed by several experimental reports.

CHAPTER I

INTRODUCTION

One of the grand challenges of bioinformatic research is to create complete computer representation of a cell *in silico*. In 2003, the human genome project successfully sequenced a particular human's genome by encoding 20,500 protein sequences from over three billion DNA base pairs in the human genome [1] and this can now be done routinely in a few weeks. Towards understanding the cellular life beyond the genome data there exist many sub-challenges including but not limited to addressing how the obtained genetic information is transferred to RNA and to ribosome where ultimately the protein will be synthesized, examining the cell membrane, microtubules and actin filaments that form the cytoskeleton, research of organelles, studying the protein-protein interfaces etc. Within the same context, another widely investigated phenomenon is the question of how protein structure and function are connected.

1.1 The complexity of protein structure and dynamics

Scientific reports dealing with protein folding mechanisms often begin with a reference to Levinthal's thought experiment on protein folding [2]. A standard illustration of this experiment involves an imaginary protein consisting of 101 amino acids. Assuming that there are only three torsional degrees of freedom defining the bond between each amino acid, the protein could exist in 3^{100} configurations. Even if the protein can sample one trillion configurations per second (i.e. 10^{20} configurations per year), it will take more than 10^{27} years to sample the whole configuration space. The fact that proteins *in vivo* fold reliably and quickly to their native conformation despite the astronomical number of possible states has been known as Levinthal's Paradox or the protein folding phenomenon. The main conclusion of Levinthal's paradigm

is that proteins do not scan all possible configurations in order to land on to their native state [3]. The thought experiment, however, remains to describe the difficulty to produce a generalized solution to the protein folding problem.

The three dimensional shape (tertiary structure) of a protein is defined by its amino acid sequence (primary structure) [4]. In addition to the amino acid sequence, the structure information of proteins depends on other external factors such as local environment [5], solvent interaction [6, 7, 8], and structural elements (i.e. α-helix, β-sheet etc) [9]. Therefore, accurate prediction of protein tertiary structure given its primary sequence can be improved by addition of such external parameters including but not limited to structural element information [10], solvent accessibility [11, 8, 12, 13], and contact number of residues [14].

The protein folding phenomenon can be divided into three major problems [15]: (i) protein structure prediction — how to predict the specific fold of a particular protein given only its primary sequence, (ii) folding speed — the kinetic question of how naturally occurring proteins fold so fast, and (iii) dynamic pathway — the thermodynamic question of how atomic interactions lead to the native fold. Although the ultimate goal is to generate an algorithm that will optimize solutions to all three problems, the vast majority of the current efforts in the area aim to solve one at a time, with much of the attention devoted to structure prediction.

The research of protein structure and folding dynamics is based on three main approaches. The comparative approach uses the information obtained from the known structures to find motifs and patterns among either similar proteins (homology modeling) or among the whole dataset ignoring any homology (protein threading). The reduced dimensionality approach includes low resolution models (e.g. as lattice proteins) and intermediate resolution models (e.g. coarse-grained bead models) to study long timescales of protein motion. Finally, the all-atom approach (e.g. molecular dynamics) provides the atomic level detail of the protein dynamics at a higher feasible

resolution. This thesis aims to contribute to the discussion of the assessment of the predicted or experimentally determined structures using a comparative data mining methodology at the residue level and to the investigation of protein unfolding motion using all-atom molecular dynamics simulations.

1.2 Structure validation by comparative modeling

1.2.1 Quality assessment and scoring functions

There is a huge gap between the number of protein sequences available and protein structures produced. Considerable effort has been put towards determining protein structures by means of both experimental (e.g. NMR, X-ray crystallography, cryo-EM) and computational techniques (e.g. comparative modeling, reduced dimensional methods (lattice models, coarse-grained bead models), all-atom analysis) Without the knowledge of the true structure of a protein, the question of how one can assess the fidelity of such determined structure to the true structure is as vexing as it is important [16, 17, 18]. A critical question facing producers of protein structures —whether it is experimentally measured or computationally predicted— is the assessment and verification of the quality of their output. All-atom *ab initio* calculations, theoretically, should be able to provide accurate prediction *in silico*; however, current computer resources are far from achieving such goal for a standard size peptide within reasonable timescales. Therefore, protein structures predicted using empirical methodologies such as comparative modeling or molecula The procedure to evaluate a produced structure is often referred to "quality assessment" or "structure validation." As will be discussed in detail throughout this chapter, a large number of assessment and validation algorithms have been proposed over the past two decades. These algorithms can be categorized by two major titles: (i) *ab initio* methods where all-atom or coarse grained energy minimization and equilibrium energy calculations are implemented, and (ii) comparative modeling which is based on the comparison to known

structures.

Homology-based comparative modeling and fold recognition is especially useful as it appears that although the number of solved structures is vast, there is only a limited set of structural motifs that most proteins adopt [19, 20, 21]. Not only is there a finite number of protein folds in nature [22, 23, 24, 25, 26] but also certain kinds of folds seem to be remarkably widespread among clearly unrelated sequences [27, 28, 29, 30, 31]. These homology based protocols often utilize previously solved structures as templates and evaluate usual and unusual patterns among them based on different metrics. The given structure is then scored by favoring usual patterns while penalizing the unusual ones. Scores produced by such algorithms are widely used to characterize the experimentally determined structures [32, 33, 34, 35, 36] as well as computationally predicted ones [37, 38]. They have also become a prerequisite for the deposition of new structures into databases such as the PDB [39].

The protocols that are developed as assessment tools mainly utilize two methodologies known as homology modeling and structure threading. For the homology modeling evaluation to produce good results, it is expected that the target protein would have a similar sequence or tertiary structure in the databank. Evaluation with structure threading, however, scans through the whole database and produces a structure based statistical score that is useful even when no close relative is detected in the template set.

Most of the threading protocols developed require initial alignment of structures generated by an alignment tool such as BLAST [40] (a simple alignment algorithm to compare primary biological information such as amino acid sequences of proteins or DNA sequences); PSI-BLAST [41] (an iterative BLAST algorithm that is used when distant evolutionary relatives are desired); CD-HIT [42, 43, 44] (an algorithm to remove redundant sequences); and HHpred/HHsearch [45] (a profile-profile comparison tool based on hidden Markov model to identify remotely related protein families).

Once the alignment and pruning process is completed, then statistical analysis is employed on the remaining data as to discover structural patterns. Over the past two decades many comparative methodologies have been developed for protein structure validation measuring different metrics within the generated template set. For example, ERRAT works by analyzing non-bonded interactions between different atom types [46], and PROVE is based on evaluating amino acid volumes [47]. The reader is referred to several excellent reviews that describe the methodologies to larger extent [48, 49, 50, 51, 52, 53, 54]. Aside from the evaluation metrics used by ERRAT and PROVE, also of particular importance to characterize secondary structures of proteins is dihedral angle analysis. The following subsection discusses the significance of dihedral angle analysis as an assessment tool.

1.2.2 Use of dihedral angles as an assessment tool

The two dihedral angles (Ramachandran angles), ϕ and ψ, describe the rotation of an amino acid around the two bonds on both sides of an α-Carbon atom (Figure 1). The collection of the two dihedral angles within a polypeptide fully defines the spatial arrangement of its backbone excluding the side chain positioning. Considerable effort has been put towards analysis of dihedral angles since Ramachandran and coworkers have proposed that there exist pairwise values that are allowed or forbidden for each amino acid [55, 56]. The favored and forbidden values are then revisited as the number of available data increased [57, 58, 59, 60, 61].

Dihedral angle analysis is widely used in protein structure validation research [62, 63, 64, 35, 36, 65, 66, 67, 68, 69]. Resulting from the high throughput investigation of protein dihedral angles, many validation tools —also called checking functions— have been proposed. Among these WHATCHECK [70] (which was developed over WHATIF [71] by Vriend and coworkers) and PROCHECK [72, 73, 32, 33, 34] (developed by Thornton and coworkers) are the two notable comparative algorithms that

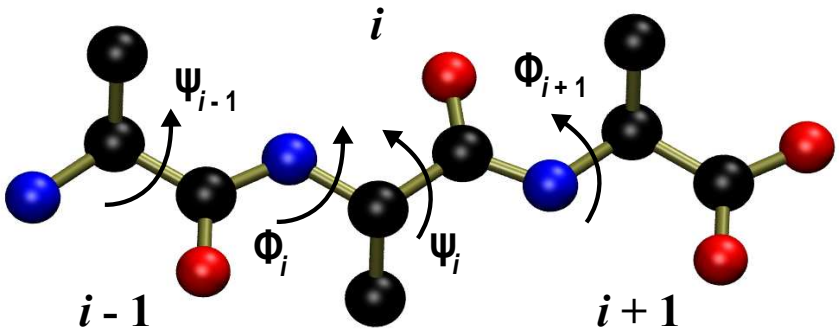

Figure 1: Schematic representation of the backbone dihedral angles in a ALA-ALA-ALA tripeptide. (Blue, Nitrogen; Black, Carbon; Red, Oxygen)

utilizes correlations in the backbone torsional angles, ϕ-ψ, in their scoring functions. However, the use of the ϕ-ψ distribution about a residue is not necessarily sufficient to describe all of the nontrivial correlations between the dihedral angles along a protein. Several groups have found that such nontrivial correlations indeed exist [74, 75, 76, 77, 78, 79, 69, 80, 81, 82, 83]. Recently, Hernandez and coworkers have introduced a novel checking function, D_2Check, which complements current assessment tools by taking ψ_i-ϕ_{i+1} angles about two adjacent residues into account in addition to ϕ_i-ψ_i angles about a single residue. The additional information assessed by D_2Check [84] is the relative correlation in the dihedral angles between two adjacent sites and their respective residue identities. In particular, an information-theory entropy [85] has been developed to gauge the degree to which these angles are statistically related to those in the training set of structures. The details and current efforts on improving the use of D_2Check is discussed in Chapter II.

1.3 Protein dynamics and folding events

Protein folding and unfolding events are fundamental events that have been quite difficult to observe *in vivo*, *in vitro*, and *in silico*. Through comparative analysis of the known structures, a common view has emerged that proteins fold mainly by the collapse of the hydrophobic core. However, hydrophobic interactions are not the

sole force that drives the folding reaction; instead, the folding pathway is driven by the sum of many different small interactions (hydrogen bonds, electrostatics, van der Waals interactions, hydrophobic interactions). As new methodologies and tools (both experimental and computational) become available, it may be possible to experimentally observe folding mechanisms to a limited extend. For example, the fast laser temperature-jump method is shown to identify transition states of fast folding proteins [86], mutational methods output ϕ and ψ values to determine amino acids that control the folding mechanism [87, 88], FRET (Förster resonance energy transfer) methods can follow closely the formation of particular contacts [89, 90], and hydrogen exchange methods are used to observe structure specific folding events [91]. The folding pathway is often considered to proceed through a series of intermediates in which structural element formations accrue [92, 93]. Multiple-state folding mechanisms are observed in folding studies of cytochrome C [94, 95], T4 lysozyme [96], staphylococcal nuclease [97], folding protein in 8m urea [98], and barnase [87]. In smaller proteins it is possible to observe two-state folding mechanisms such as in chymotrypsin inhibitor 2 (CI2) [99, 100].

Quantitative interpretation of the thermodynamics and kinetics of protein folding and unfolding events in terms of a statistical energy landscape approach has been introduced by Bryngelson and Wolynes in the late 1980s [101, 102, 103, 104, 105]. Statistical characterization of these events in terms of energy distribution and other measurables has then been studied extensively on many systems such as minimalist lattice models [106, 107, 108, 109, 110, 111, 112, 113], coarse grained models [114, 115], and finally all-atom models [116, 117, 118, 118]. Figure 2 displays an illustrative example of a representative folding energy landscape with a stable intermediate and multiple transition states (the main graph) and no intermediate —two-state folding— and a single transition state (the inset). The complete funnel of most protein folding processes, however, exhibits a more complex landscape (see Figure 1 in Wolynes *et al*

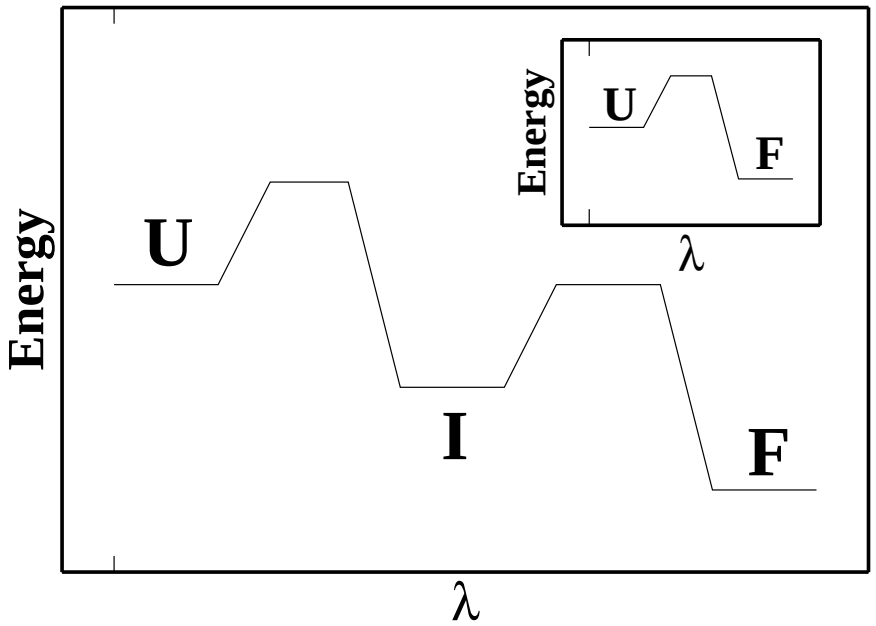

Figure 2: Schematic representation of the free energy profile along the folding pathway, λ, (i) with possible low energy intermediates (main graph), and (ii) with no intermediate present (inset). (U, unfolded state; I, intermediate states; F, native fold)

[105]). Real proteins, as proposed by Levinthal, do not visit all possible energy configurations to find the native fold. Instead, they follow a folding pathway driven by a reasonable energy bias (such as hydrophobic interactions or electrostatics) against locally unfavorable configurations [119]. Such bias towards the lowest energy configuration over the remainder of the effective energy surface will decrease the search space significantly by reducing the number of pathways that need to be postulated [120]. There have been many biased methodologies developed to achieve fast sampling of the folding energy landscapes *in silico*. Next, several of those methodologies are discussed within the framework of all-atom molecular dynamics simulations.

1.3.1 Protein unfolding dynamics at elevated temperatures

Applying Newton's laws of motion (MD) [121, 122, 123, 124] to every atom in a solvated system, even when using empirical force-fields, requires a massive amount of computational time. Efforts to simulate protein folding in the folding direction are

restricted by current resources, since such simulations should be very long (generally from microseconds to milliseconds) to be able to observe a single folding event. At this point, simulating unfolding events instead of folding events has found strong interest for several reasons. First, sampling limitations are greatly reduced because the unfolding simulation has to start from the best-defined state (the native state) compared to folding simulation which has to start from the least characterized state (the denaturated state). Secondly, the chance to observe unfolding events increases at elevated temperatures. Most proteins unfold in less than 1 ns at temperatures higher than 498 K.

The unfolding pathway observed at elevated temperature is claimed to be identical to the path that should be observed at standard temperatures. For proteins that have two-state energy landscape, the transition state (top of the plateau in the inset of Figure 2) is likely to be the same. Itzhaki *et al.* explained that with the principle of microscopic reversibility in their study comparing the folding and unfolding kinetic behavior displayed by the 64-residue protein, chymotrypsin inhibitor 2 [125]. Temperature independence of protein folding pathways are also proposed in a recent β-heptapeptide study by van Gunsteren and co-workers [126]. They have observed common intermediates in the MD trajectories they gathered at four different temperatures (i.e. 298/340/350/360 K). Similar findings have been reported on lattice model simulations by Dinner and Karplus [112]. Finally, Daggett and coworkers have recently showed that this reversibility is also valid along multistate unfolding routes of large-scale proteins and confirmed that increasing temperature accelerates protein unfolding without changing the pathway of unfolding [127]. 498 K is well above any natural biological temperature, and is also above the protein melting temperature, T_m. So an experimental system under these conditions would exhibit different dynamics than the biological case. The water system in the computer model, however, remains as a metastable and superheated liquid because neither chemical bond

breaking-and-making or evaporation pathways are available to it. The key assumption is that the dynamical pathways also remain in the same universality class, and thus to confirm the predictions of correlation functions using temperature acceleration additional tests are required. One way to confirm the pathway observed via high temperature molecular dynamics simulations is to apply biasing forces, along the proposed pathway, that will eventually compensate for the free energy gradient. The next section discusses several of the biasing methodologies.

1.3.2 Biased molecular dynamics

Besides temperature acceleration, the efficiency of searching the energy landscape of protein folding/unfolding events can also be increased by the addition of different external biases.

Replica exchange molecular dynamics or parallel tempering [128] overcomes the barriers confronted along the energy landscape by exchanging non-interacting replicas that are obtained at several temperatures. The method allows the calculation of thermodynamic observables as a function of temperature by generating the canonical ensemble probability distribution. The probability distribution is created by using weighted-histogram analysis techniques. Sugita and Okamoto has demonstrated the efficacy of the method on the folding study of penta-peptide Met-enkephalin [128]. Replicas do not necessarily have to be exchanged according to temperature sampling. Recently, the same group has shown that besides temperature other parameters of the potential energy can be used as the exchange parameter [129]. Okur *et al* showed that energy convergence is enhanced when the system is coupled to a high temperature structure reservoir [130]

Adaptive biasing force molecular dynamics [131] calculates the mean force, along the selected reaction coordinate, which is later canceled out by an equal and opposing biasing force. The biasing force is computed from the derivation of the free energy

— hence the name adaptive. The applied force pushes the system to escape from local minima and thus sample the whole conformation space even when the surface is rough. The net force applied along the reaction coordinate is the fluctuating part of the instantaneous force. Therefore, the process is almost the same as a random walk with zero mean force. Adaptive biasing force MD has shown to produce good results when the reaction coordinate is not coupled to any other degrees of freedom. When the said decoupling is not satisfied, the energy landscape may be disturbed by strong initial fluctuations which, in fact, can be solved by accumulating large number of force samples before starting to estimate the biasing force. The computation of the biasing force can be done by deriving the free energy with respect to cartesian coordinates only [131] or with respect to cartesian coordinates and time [132]. The latter was tested in a study of N-acetylalanyl-N'-methylamide to construct free energy as a function of the backbone dihedral angles, ϕ and ψ. They have demonstrated the use of the new derivation on the N-acetylalanyl-N'-methylamide and estimated free energy as a function on the ψ and ϕ dihedral angles.

One other way that reduces the cost to sample the energy landscape is steered molecular dynamics. Steered molecular dynamics (SMD) is often utilized in conjunction with Jarzynski's nonequilibrium work relation. Next, a brief introduction on the basic theory and current applications of steered molecular dynamics and Jarzynski's relation is provided. More information including a proof on classical Hamiltonian systems is discussed in Chapter III.

1.3.3 Steered molecular dynamics and Jarzynski's inequality

The domain of the energy landscape of most proteins, even for small peptides, has a high dimensionality. The identification of an unfolding pathway is therefore useful because it greatly reduces this dimensionality. Once identified, the energetics along this

pathway is determined by the so-called potential of mean force (PMF). [See e.g. reference [133].] The importance of the PMF as well as the difficulty in calculating it has led to the development of far too many approaches to list here. Instead, we focus on those approaches which rely on sampling the states directly from trajectories. Unfortunately, the use of unconstrained trajectories is cost prohibitive when the processes of interest are very slow and dominated by deep minima. Instead, SMD can accelerate such processes by applying steering forces along the chosen unfolding pathway. Such a non-equilibrium process would not seem to provide the unconstrained structures required to obtain the equilibrium PMF. This problem was resolved by Jarzynski when he showed that an appropriately weighted average of the non-equilibrium work over many such SMD trajectories leads to the PMF [134, 135]. Jarzynski's relation (also referred as "Jarzyinski's inequality" or "Jarzynski's equality") connects non-equilbirium processes to equilibrium properties (i.e. the free energy). The free energy difference $\Delta G_{\xi' \leftarrow \xi}$ to take a system from the state ξ to a new state ξ' is provided by Jarzynski's equality,[134, 135]

$$\Delta G_{\xi' \leftarrow \xi} = -\frac{1}{\beta} \ln \left\langle e^{-\beta W_{\xi' \leftarrow \xi}} \right\rangle \ , \tag{1}$$

where $W_{\xi' \leftarrow \xi}$ is the work along a nonequilibrium trajectory from ξ to ξ', and the average is over many trajectories starting with the equilibrium configuration at ξ such that $\left\langle e^{-\beta W_{\xi' \leftarrow \xi}} \right\rangle = \int dW_{\xi' \leftarrow \xi} e^{-\beta W_{\xi' \leftarrow \xi}} P(W_{\xi' \leftarrow \xi})$ where $P(W)$ is the work distribution.

Over the past decade since Jarzynski first introduced his eponymous inequality, significant amount of work has been published discussing various aspects of the original theorem including but not limited to derivations for specific systems. Jarzynski's equality has been validated numerically on several systems such as deca-alanine stretching by Park and Schulten [136], Ace-Alanine$_8$-NMe unfolding and ligand diffusion in globins by Xiong *et al* [137], and Angeli's salt decomposition by Torras *et al* [138]. It has been compared to existing biased MD techniques, such as to umbrella sampling [139] and to targeted MD [140] yielding comparable results. It has also been

verified in the context of experimental results such as RNA unfolding by Liphardt *et al* [141] and a mechanical oscillator [142].

In a typical SMD study, the system is first driven away from equilibrium by superimposing time-dependent harmonic constraints along the reaction coordinate (ξ) in a series of simulations so that the hamiltonian is biased with the addition of the following harmonic potential:

$$h_\lambda(r) = \frac{k}{2}(\xi(r) - \lambda)^2 \tag{2}$$

The work done on the system is calculated for each simulation to acquire the work distribution. The work distribution is then averaged using Jarzynski's equality

$$G(\xi_t) = G(\xi_0) - \frac{1}{\beta} \ln \left\langle e^{-\beta W_{\xi_t \leftarrow \xi_0}} \right\rangle_0 \ , \tag{3}$$

to calculate the PMF along the reaction coordinate.

Although SMD, compared to unconstrained MD, effectively reduces the processing cost in modeling large conformational changes of biomolecular systems, the amount of force applied (ranging typically from 500 pN to several thousands pN) is far larger than that applied in AFM experiments (up to a couple hundred pN) from which SMD aims to reproduce the results. As the applied force (thus work) reads larger values, the distribution of work gets distorted from Gaussian behavior which ultimately causes the calculated PMF to be dominated by the lowest energy trajectories. In order to overcome this shortcoming, three solutions may be proposed: creating a large enough ensemble (as large as millions), lowering the pulling velocity (as low as the reversible speed), or equilibrating the system at short intervals. All of these solutions will again increase the computational cost dramatically. Recently, adaptive SMD methodology [143] has been introduced to restrain the ensemble work distribution within a Gaussian nature at all points along the reaction coordinate (e.g. the unfolding of neuropeptide Y). The adaptive SMD methodology is discussed along with a heuristic proof of the applicability of Jarzynski's equality on adaptive integrations in Chapter III.

1.4 Organization of Thesis

Aside from Chapter I —introduction and background— and Chapter VII —concluding remarks and scientific significance— the thesis contains five chapters that are adapted from previously published or recently submitted articles.

Chapter II describes the framework of a novel checking function, D_2, which has been recently developed by our group [84]. D_2 is basically a measure of dihedral-angle information entropy that compares the $\phi - \psi$ angles about a single residue and $\psi - \phi$ angle pairs about two adjacent residues to the library of the corresponding angles obtained from the structures in the protein data bank. Taking $\psi - \phi$ pairwise angles into account provides 400 possible distributions in addition to 20 possible Ramachandran angle distributions. The additional data obtained from the $\psi - \phi$ angle pairs about two adjacent residues introduces complementary stereochemical information to standard Ramachandran analysis and increases the awareness of the effect of nearest-neighbor frequency on the pairwise dihedral angle distributions. The work published in 2004 has demonstrated that the D_2 distribution of the structures in the protein data bank is a Gaussian centered about 0 with 99.4% of all structures covered in the range of $|D_2| < 3$. The produced distribution means that D_2 can successfully identify protein structures whose angles are mostly atypical with respect to the 420 correlated angle distributions. The D_2 methodology can also be reduced to residue level analysis to identify those atypical angles within a given structure. Residue level analysis has shown that only if a significant number of residues of a given protein structure reads atypical angles then the absolute value of the overall D_2 value becomes greater than 3. Residue level D_2 compatibility of a given protein structure is displayed in a compact form using a *color strip*. The color strips visualize the structure of a given protein as to observe the intensity of atypical structures within the protein of interest. The work is further extended towards analyzing structural similarities/differences amongst protein families and also investigating structural effects of

mutations. Graphical representation of the *color strip difference* has been found to be quite effective in characterizing a family of tested mutants (staphylococcal nuclease).

Chapter III addresses the adaptive steered molecular dynamics methodology in detail. A review and a direct proof of Jarzynski's equality is followed by a discussion of its implementation in molecular dynamics simulations — steered molecular dynamics. Steered molecular dynamics is reported to yield good results in a number of previous studies [136, 137, 138, 139, 140, 141, 142]. However, as will be discussed in Chapter III, there are some major drawbacks that one can observe when using steered molecular dynamics. One such drawback occurs in simulating large systems which require large amount of work to steer the system along a preselected path. In such cases, one must either apply smaller forces at the expense of increasing the overall simulation time or create a massive number of ensemble for an accurate estimate of the free energy. At this point, adaptive steered molecular dynamics [143] does a good job restraining the applied force within a narrow range so that the work distribution is always Gaussian.

In Chapter IV the use of adaptive steered molecular dynamics simulations to study the helix coil transition of deca-alanine is presented. In vacuum the free energy landscape for the stretching of deca-alanine was previously calculated by Schulten and coworkers [136]. Adaptive steered molecular dynamics [143] are utilized to reproduce the free energy profile along a trivial single dimension stretching pathway. The adaptive SMD methodology requires significantly less computing resources to estimate the PMF compared to the standard SMD technique. In addition to the analysis in vacuum, the helix coil transition of deca-alanine has been investigated in explicit water solvent. The PMF of the solvated system exhibits a narrower energy landscape with a rather flat plateau towards the coil formation. This behavior is explained by the stabilizing effect of water molecules on the peptide. Water molecules are observed to rush in to replace broken intra-peptide hydrogen bonds to stabilize the energetically unstable intermediates formed as the peptide starts to lose hydrogen bonds.

Finally, Chapter V and Chapter VI discuss the use of steered molecular dynamics on a rather large system, i.e. the unfolding dynamics of a 36 residue neuropeptide Y and various mutants of peptide YY. Neuropeptide Y (NPY) and porcine peptide YY (PYY) adopt a common stable pp-fold which consists of an α helix and a poly-proline tail that interacts via side chains. The unfolding pathway of these neurotransmitters is not as trivial as the stretching of deca-alanine due to these hydrophobic side chain interactions. At the biologic body temperature, 310 K, NPY did not unfold in unconstrained MD simulations. Temperature acceleration has been employed to characterize the unfolding mechanism. At 500 K, which is well above the melting temperature of proteins, NPY is observed to lose the stabilizing contacts within the first couple hundred ps window. As described in Section 5.2.1, the unfolding mechanism observed at a temperature this high, sometimes leads to experimental errors so that it projects an inaccurate mechanism compared to low temperature unfolding. In order to eliminate possible skepticism, the proposed unfolding pathway is tested using both standard SMD and adaptive SMD methodologies. Results obtained from the standard SMD simulations produced a PMF that fails to represent the whole sample space in the sense that it is always dominated by the lowest energy trajectories. This is much expected in the case of NPY unfolding since the energy required to break strong side chain interactions quickly lead to values of 10 $k_{\mathrm{B}}T$. When dealing with these high values, the Gaussian nature of the energy distribution tends to get distorted unless a massive number of events are sampled. Results obtained from the adaptive SMD simulations showed that the work distribution is Gaussian at all points along the unfolding pathway. Therefore, the PMF always represents a sufficient amount of the trajectories and thus is more accurate. Within the same context, porcine peptide YY (PYY), which is in the same family as NPY and share the common pp-fold, and several of its single residue mutants are also investigated through adaptive SMD

calculations. Recently, Waegele and Gai reported a temperature-jump infrared spectroscopy study to acquire quantitative values of folding and folding rates of PYY and the mutants. Adaptive SMD simulations of the unfolding of PYY and the mutants are implemented along the same unfolding pathway as NPY. Since, PYY adopts the same pp-fold as NPY adaptive SMD is applied along the same unfolding mechanism that was observed for the neuropeptide Y. The preliminary results, however, do not produce well characterized PMFs as observed in the study of NPY unfolding. Possible explanations to this behavior and plausible solutions within the framework of adaptive SMD methodology are discussed as a future direction.

CHAPTER II

SECONDARY STRUCTURE PROPENSITY OF PROTEINS IN TERMS OF RESIDUAL D_2 SCORE ANALYSIS

2.1 Overview

Recently, a new and practical checking function of dihedral angle correlation has been developed by our group to to assess secondary structure propensity in proteins [84]. The checking function D_2 assesses not just the Ramachandran angle pairs, but also the ψ_i-ϕ_{i+1} angle pairs between adjacent residues that had heretofore been largely ignored by other checking functions. This new checking function is not directly correlated to structural fidelity, but it does signal deviations of a protein from the most probable experimentally derived structures in the PDB at the protein scale. In some cases, such deviations are markers of incorrect structures, but in others they signal unusual propensities due to other factors.

This chapter discusses the theory, use and several examples of the D_2 checking function at both protein level and residue level. First a review of the D_2Check is discussed in terms of a short proof. At this point, the ΔS and D_2 distributions have been reproduced from a larger data set compare to the data set used in the 2004 paper. Then, D_2Check at the residue level is introduced and how it can be used to identify atypical structures by means of a graphical representation is addressed. After that, the potential application of the residue level D_2 methodology on the structural comparison of mutants is described (e.g. staphylococcal nuclease is compared to five of its mutants in terms of conformational analysis). And finally, the D_2Check web server is presented.

2.2 D_2Check as a complementary validation tool

In previous work [84] an information entropy, $S(\vec{q})$, was developed based on the information to be found in the correlation in the dihedral angles about each residue and between adjoining residue pairs given some specified structure $\vec{q}$. It relies on a knowledge of the probability distributions for the sets of all ϕ_i-ψ_i [55, 56] and ψ_i-ϕ_{i+1} pairs of a protein that had earlier been calculated by several others [144, 145, 146, 147] and was updated using the October 2004 release of the PDB [84]. The entropy difference, ΔS, between the standard information entropy $S^{\circ}(\vec{q})$ —in which each ϕ-ψ and ψ-ϕ angles assumes the most probable angles— and the information entropy $\Delta S(\vec{q})$ was calculated across the $2,762$ nonredundant experimentally derived protein structures in the PDB at the 90% level of sequence identity. These values were used to construct a probability distribution for a given protein to have a value of ΔS. The distribution is updated as of February 2011 to include the structures that were released since October 2004. Figure 3 evaluates both 2004 and 2011 PDB entries and shows that as the number of structures analyzed increase the distribution maintains the same width but shifts towards lower values.

The symmetric behavior of the ΔS distribution suggested the use of a checking function, D_2, which can be used to assess the propensity for secondary-structure correlation contained within a given protein in comparison with those in the reference set of structures. The analytic forms for all of these quantities are reviewed below. The extension of these quantities to an information entropy and associated meastures at the residue level subsetuently forms the bulk of this section and is the central result of this work.

2.3 D_2Check at the protein level

The domain of the information entropy function for a given protein structure, can be defined in terms of the the dihedral angle pairs grouped within a single sequence of

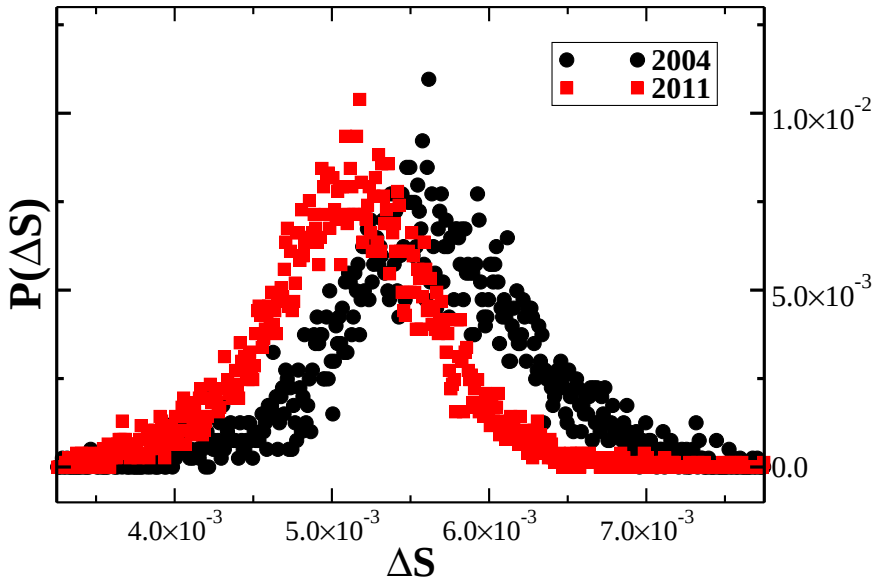

Figure 3: The distribution of ΔS values of the structures in the protein data bank as of February 2011 (red) is compared to the distribution of ΔS values of the structures in the protein data bank as of October 2004 (black).

length $(2n-3)$ as

$$\Upsilon_{2i-1} \equiv (\psi_i, \phi_{i+1}) \qquad \text{for } 1 \leq i \leq n-1 \tag{4a}$$

$$\Upsilon_{2i} \equiv (\phi_{i+1}, \psi_{i+1}) \qquad \text{for } 1 \leq i \leq n-2\,, \tag{4b}$$

in which the k label alternates between the dihedral angles refering to a single residue and those refering to an adjoining residue pair. The labels of the residues corresponding to this sequence can be grouped in a similar fasion as

$$\xi_{2i-1} \equiv (R_i, R_{i+1}) \qquad \text{for } 1 \leq i \leq n-1 \tag{5a}$$

$$\xi_{2i} \equiv R_{i+1} \qquad \text{for } 1 \leq i \leq n-2\,. \tag{5b}$$

The information entropy of a given structure, $\vec{q}$, is defined as

$$S(\vec{q}) = -\sum_{k=1}^{2n-3} P_{\xi_k}(\Upsilon_k(\vec{q})) \ln P_{\xi_k}(\Upsilon_k(\vec{q}))\,. \tag{6}$$

It gave rise to a maximum entropy,

$$S^{\circ}(\vec{q}) = -\sum_{k=1}^{2n-3} \bar{P}_{\xi_k} \ln \bar{P}_{\xi_k} \tag{7}$$

in which the maximal values are defined as

$$\bar{P}_{\xi_k}(\vec{q}) \equiv \max_{\Upsilon_k} P_{\xi_k(\vec{q})}(\Upsilon_k) \; . \tag{8}$$

The difference between Equations (7) and (6), normalized by the number of entries, gave rise to an entropy difference:

$$\Delta S(\vec{q}) = (S^{\circ}(\vec{q}) - S(\vec{q}))/(2n-3) \; . \tag{9}$$

The values of ΔS across a given database can subsequently be collected and its statistics can be summarized by a protein-level distribution function, $P_{\rm prot}(\Delta S)$.

The D_2 checking function was subsequently defined as a gauge of the likelihood of the particular value of $\Delta S(\vec{q})$ such that a value of 0 corresponds to the most likely value, and increments correspond to standard deviations from this maximal value. Specifically,

$$D_2(\vec{q}) \equiv \begin{cases} \sqrt{2}{\rm erf}^{-1}(2I-1) & \text{if } \Delta S < \overline{\Delta S} \\ \sqrt{2}{\rm erf}^{-1}(1-2I) & \text{if } \Delta S \geq \overline{\Delta S} \end{cases} \; . \tag{10a}$$

where the probability integral is

$$I(\vec{q}) = \begin{cases} \int_0^{\Delta S(\vec{q})} P_{\rm prot}(\Delta')\, d\Delta' & \text{if } \Delta S(\vec{q}) < \overline{\Delta S} \\ \int_{\Delta S(\vec{q})}^{\infty} P_{\rm prot}(\Delta')\, d\Delta' & \text{if } \Delta S(\vec{q}) \geq \overline{\Delta S} \end{cases} \; . \tag{10b}$$

Although these expressions are quite formal, they are easy to compute numerically. The use of Equation (10b) effectively uniformizes the distribution so that D_2 displays Gaussian statistics across the reference set of structures.

D_2 values appear to summarize the degree of secondary structure propensity with respect to the positions of the dihedral angles in a given protein structure. Figure 4 displays the Gaussian behavior of D_2 distribution based on PDB entries as of February 2011. For comparison purposes, the distribution based on older data set (2004) is also depicted on the same graph.

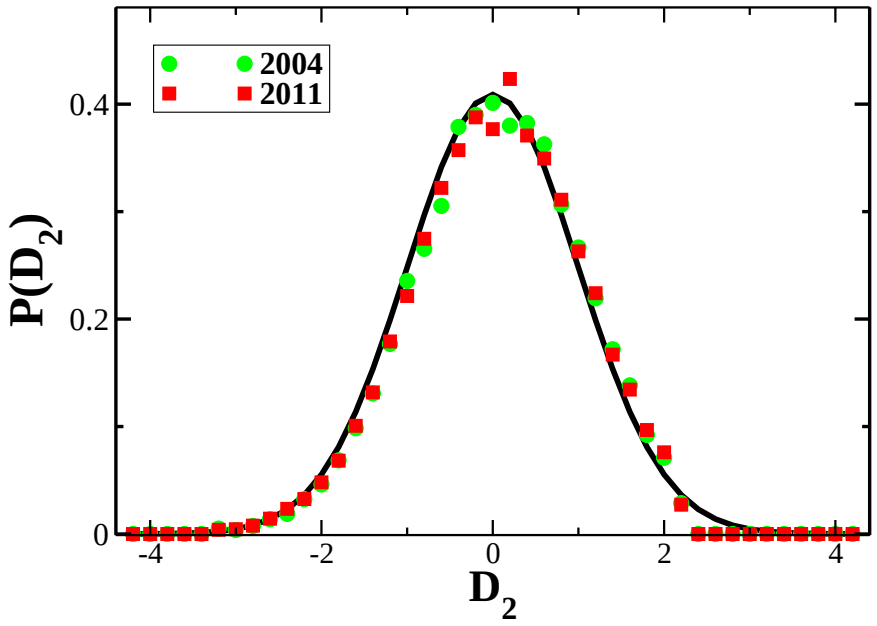

Figure 4: The distribution of D_2 scores of the structures in the protein data bank as of February 2011 (red) is compared to the distribution of D_2 scores of the structures in the protein data bank as of October 2004 (black).

2.4 D_2 Check at the residue level

The additive structure in Equations (8) and (6) also suggests that ΔS can be extended to a partial information entropy for each dihderal angle pair, k. That is, the information entropy at the residue level can be written as

$$S(\Upsilon_k(\vec{q})) = -P_{\xi_k}(\Upsilon_k) \ln P_{\xi_k}(\Upsilon_k) \ , \tag{11}$$

with a corresponding standard entropy,

$$S^{\circ}(\Upsilon_k(\vec{q})) = -\bar{P}_{\xi_k} \ln \bar{P}_{\xi_k} \ . \tag{12}$$

The entropy difference at the residue level follows readily from the difference between these last two equations:

$$\Delta S(\Upsilon_k(\vec{q})) = S^{\circ}(\Upsilon_k(\vec{q})) - S(\Upsilon_k(\vec{q})) \ . \tag{13}$$

It should be apparent that these definitions allow one to rewrite the entropy difference at the protein level in terms of those at the residue level:

$$\Delta S(\vec{q}) = \frac{1}{2n-3} \sum_k \Delta S(\Upsilon_k(\vec{q})) \ , \tag{14}$$

or even more compactly as an average of the residue values,

$$\Delta S(\vec{q}) = \langle \Delta S(\Upsilon_k(\vec{q})) \rangle_k \ , \tag{15}$$

in which the subscript k on the angle brackets is used to indicate that it is the index of the domain for the average.

The probability distribution, $P_{\rm prot}(\Delta S)$ of ΔS across the entire database was obtained in the previous section. In principle, corresponding residue-level distributions of the entropy differences can be obtained: $P_R(\Delta S)$ labeled by a single residue type in the case of the dihedral angles around a particular residue R, and $P_{R,R'}(\Delta S)$ labeld by a residue pair in the case of the dihedral angles between a given pair of adjoining residues R and R'. These 420 distributions could then be used to obtain the D_2 checking function at the residue level in direct analogy to prior work. Such an approach, however, has two severe disadvantages: *(i)* it involves substantial data gathering that may or may not reveal significantly distinct distributions while possibly suffering from a lack of data for particular sets, and *(ii)* such residue-level checking functions would not necessarily connect directly to the values of the protein-level checking functions. However the protein-level distribution, $P_{\rm prot}(\Delta S)$, was seen to be nearly Gaussian, and this is certainly suggestive that the residue-level distributions, $P_R(\Delta S)$ and $P_{R,R'}(\Delta S)$, are Gaussian (or nearly so). That is, if the latter were true, then certainly the former would also be true because of the properties of Gaussians. The converse does not necessarily follow, and the component Gaussian distributions may also not necessarily share the same width —viz σ— as the combined distribution. But if one replaces the residue-level distributions with the protein level distribution in surmising the probability of a given $\Delta S(\Upsilon_k(\vec{q}))$, then all of these converse relations

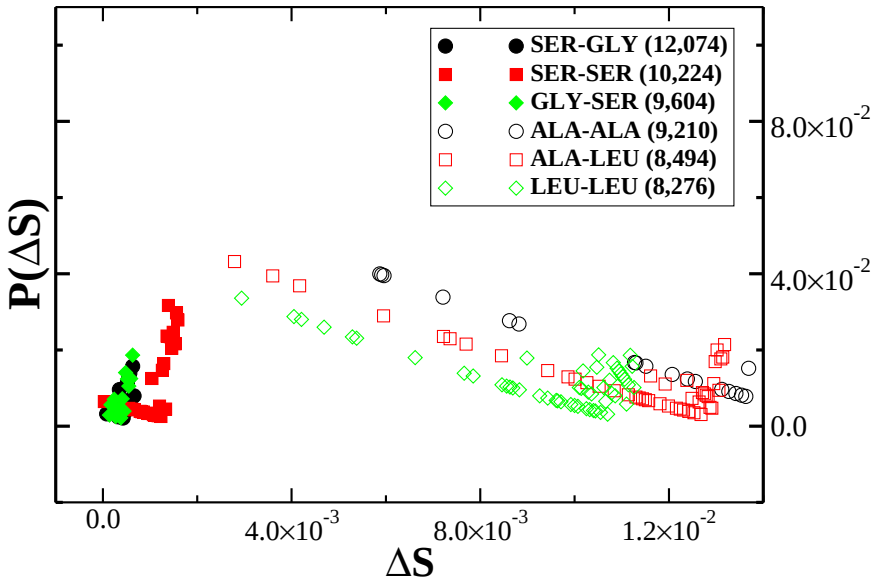

Figure 5: The distribution of ΔS values at the residue level for the most populated residue pairs: SER-GLY (12,074), SER-SER (10,224), GLY-SER (9,604) ALA-ALA (9,210), ALA-LEU (8,494) and LEU-LEU (8,276). The number of observations of the corresponding pairs are given in parentheses.

do follow. That is, the collective statistics of the components at the residue level will be in agreement with that of the sum. Figure 5 shows the residue level ΔS distributions of SER-GLY (12,074), SER-SER (10,224), GLY-SER (9,604) ALA-ALA (9,210), ALA-LEU (8,494) and LEU-LEU (8,276). These residue pairs are the most populated (counts are in parantheses) pairs among the 400 possible amino acid pairs within the NR90 database. It should be clear that even so, there is considerable noise in the ΔS distributions at the residue level compared to the protein level ΔS distribution displayed in Figure 3 and that they are quite sensitive to the pair identity. This is generally in keeping with the foreshadowed statement *(i)* above, and confirms that there isn't yet enough data to construct residue-level ΔS distributions.

Qualitatively, the graphs do suggest that there are peaks in these distributions and that they are not far from those of the protein-level ΔS distribution. Thus, we use the protein-level distribution $P_{\rm prot}(\Delta S)$ to construct the residue level D_2 checking function, with the added advantage that it must therefore necessarily satisfy statement

(ii) above. Specifically, $D_2(\Upsilon_k(\vec{q}))$ is obtained by direct substitution of $\Delta S(\Upsilon_k(\vec{q}))$ in Equation (10).

2.4.1 Color strip and color strip difference

With the definition of $D_2(\Upsilon_k(\vec{q}))$, there are now $2n-3$ values for any given protein structure in addition to the protein-level D_2 value. Although there are many ways in which this can be visualized, such as in the scatter plots of Figure 4, it is helpful to have a compact form that can readily be compared between different proteins. For this purpose, a color strip has been developed in which the values of the residue-level D_2 scores are represented using a specified color range. Specifically, red, green and blue colors correspond to -3, 0 and $+3$ of the D_2 values, respectively, with other values consisting of a sum of these colors. Color strip difference

2.5 Results and discussion

2.5.1 D_2Check at the protein level

As of February 2011, there are 167,907 chains in 71,264 entries in the protein data bank. Instead of using the complete data set, a sub-library is created at 90% redundancy level. The resulting library consists of 32,016 non-redundant structures. From this library, the structures with missing residues or atoms are also removed for obvious accuracy purposes. Structures that contain 19 or less number of residues are also eliminated. The final library (called NR90) had 7,699 structures compared to 4,013 used in 2004. Figures 3 and 4 show the ΔS and D_2, respectively. As seen in Figure 4, a near Gaussian distribution about 0 is observed for the D_2 values. Most of the 7,699 structures fall in the D_2 range of ± 3. Only 17 out of the 7,699 structures fall outside of this range which indicates a good characterization of the protein data bank in terms of ϕ_i-ψ_i and ψ_i-ϕ_{i+1} analysis. As far as the deposited theoretical models considered, only 3 out of 1,032 structures have $_2$ values smaller than -3 or greater than +3. Table 1 lists these 20 structures and corresponding chain identifiers, D_2 and

Table 1: There are 17 experimental (top section) and 3 theoretical (bottom section) structures with absolute D_2 greater than 0

PDB ID	D_2	D_2'	# of $\Delta S = 0$	(# of residues)
1t6fB	-3.19	-2.65	8	(36)
1h2sB	-3.12	-2.06	20	(59)
1piqA	-4.41	-3.06	9	(30)
1n7sA	-3.06	-2.69	13	(62)
1n7sB	-3.09	-2.74	14	(67)
3ahaB	-4.41	-2.93	9	(32)
1n7sC	-3.32	-3.12	16	(78)
2k9jB	-3.32	-2.51	18	(42)
1aikC	-3.04	-2.34	8	(33)
3ahaA	-3.16	-2.16	11	(34)
1l2pA	-3.29	-2.59	18	(60)
2guvD	-3.32	-2.91	13	(55)
1jekA	-3.01	-2.82	4	(39)
1psmA	4.41	4.41	0	(37)
1jekB	-3.24	-2.69	9	(33)
2akfC	-3.32	-2.81	9	(31)
2jo4D	4.41	4.41	0	(19)
1llkA	-3.06	-2.71	49	(260)
1z2hA	-3.19	-2.56	43	(152)
1sewF	-4.41	-4.41	6	(23)

D_2' values, number of residues with $\Delta S = 0$ and number of residues in the structure.

D_2' is a dummy measurable introduced to explore the effect of residual D_2 scores on the overall D_2 score of a given structure. It is simply calculated by excluding the ψ-ϕ and ϕ-ψ pairs whose ΔS is close to zero. For example, for the first entry listed on Table 1 —1t6fB— the overall D_2 is -3.2. If we removed the ψ-ϕ and ϕ-ψ pairs that have $\Delta S = 0$ (i.e. 8 of the 35 pairs), then the new overall D_2 (i.e. D_2') is -2.65. It was noticed that the absolute D_2' value for some atypical structures whose absolute D_2 value is greater than 3 is now smaller than 3. Among the structures listed in Table 1, this behavior is observed for 11 of the 17 experimental structures and 2 of the 3 theoretical models. This observation indicates that the origin of atypicality for most of the structures is that considerable number of residues and residue pairs have

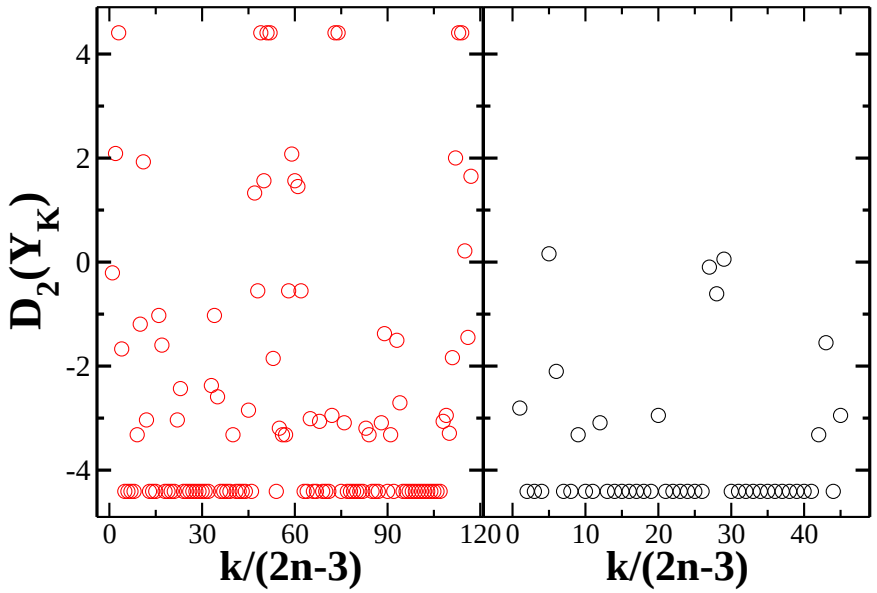

Figure 6: The residual D_2 scores of 1h2sB (red) and 1sewF are displayed on the left (red) and on the right (black), respectively. Many residue pairs, of both 1h2sB and 1sewF, appear to have D_2 scores greater than 3 or lower that -3.

the most probable dihedral angles of the corresponding dihedral angle distributions.

2.5.2 D_2Check at the residue level

For a given structure, individual residue level D_2 scores can be illustrated in several ways. For example, Figure 6 displays them as a scatter graph for two of the atypical structures: 1h2sB and 1sewF. This however is not a good representation since the quantitative comparison of the number of typical and atypical residues require a closer look and usually counting the data points. To provide a simpler illustration, which allows the user to qualitatively measure the number of atypical residues at a glance, a linear, color coded representation —color strip— has been developed. A color strip is defined by a specified three-color scheme in which the values of the residue-level D_2 scores -3, 0 and $+3$ of the residual D_2 values correspond to red, green, and blue, respectively. If the color strip generated for a given protein is rich in color red or blue, it means that the structure is composed of many residues that have atypical dihedral angle arrangements and therefore probably folded onto a tertiary structure which is

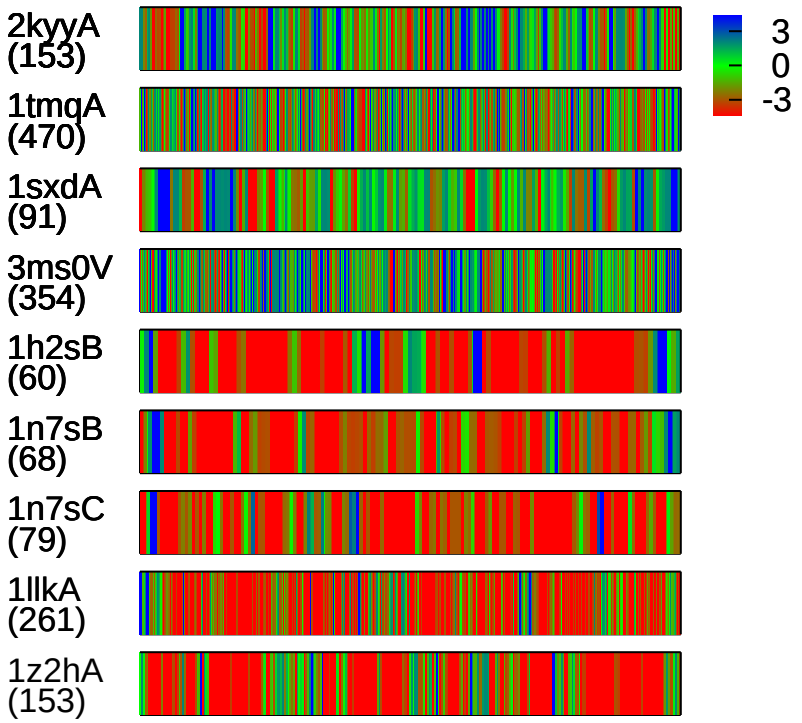

Figure 7: Examples of color strip representations: Top four —2kyyA, 1tmqA, 1sxdA and 3ms0V— are experimentally determined structures all having protein level D_2 scores close to 0; middle three —1h2sB, 1n7sB and 1n7sC— are experimentally determined structures and bottom two —1llkA and 1z2hA— are theoretical models all having protein level D_2 scores smaller than -3. In parentheses number of amino acids belonging to corresponding chain is given. Color coding is given on the top right as blue, D_2 score close to 3; red, D_2 score close to -3; green, D_2 score close to 0.

not commonly seen within the protein data bank.

Figure 7 demonstrates how effective color strips are to characterize typical and atypical structures at a glance. 2kyyA, 1tmqA, 1sxdA and 3ms0V appear to be rich in green color meaning that most of their residues have absolute D_2 values smaller than 3. On the other hand, 1h2sB, 1n7sB, 1n7sC, 1llkA and 1z2hA appear to be rich in red and blue colors meaning that most residues in those chains have absolute D_2 values larger than 3.

2.5.3 Color strip difference analysis of staphylococcal nuclease mutants

The difference between the color strips of two structures of the same length can be represented in a similar scheme called color strip difference. Since it measures the difference between two color strips, the color coding should range from +6 to -6. In order to differentiate between color strip and color strip difference, different colors are assigned to +6, blue; -6, red; 0, white.

By definition, the color strip difference is only applicable to structures of the same length. This makes the color strip difference analysis a good candidate for the investigation of structural implications of mutations, or characterization of decoys. As an example study, the analysis is applied to study the structural differences between the wild type staphylococcal nuclease and five of its mutants.

Color strip difference analysis successfully identifies the structural changes around the mutation site with respect to the native staphylococcal nuclease. This is expected since due to the *cis* and *trans* conformations of the LYS116-PRO117 neighbors, the peptide bond between the two result in different conformations of the 112-117 loop. The conformational changes of the 112-117 loop upon mutation of LYS116 has been studied extensively [148, 149]. Since the changes is due to the torsional deformations, the identification of these deformations by the D_2 scoring scheme is already expected. Color strip difference comparison of the staphylococcal nuclease to its mutants, has identified two more key regions that undergo significant conformational changes. These regions interestingly point towards two of the three α-helices of the staphylococcal nuclease as shown in the cartoon in Figure 9. Residues 49-55 and 91-95 (marked as a and b, respectively in both Figures 8 and 9) are reported to be potential candidates for binding interfaces as it is the case for the 112-117 loop. Although, α-helices are mostly resistive to torsional deformation, the single point mutation in the active site of staphylococcal nuclease appears to cause unexpected conformational changes at both helices.

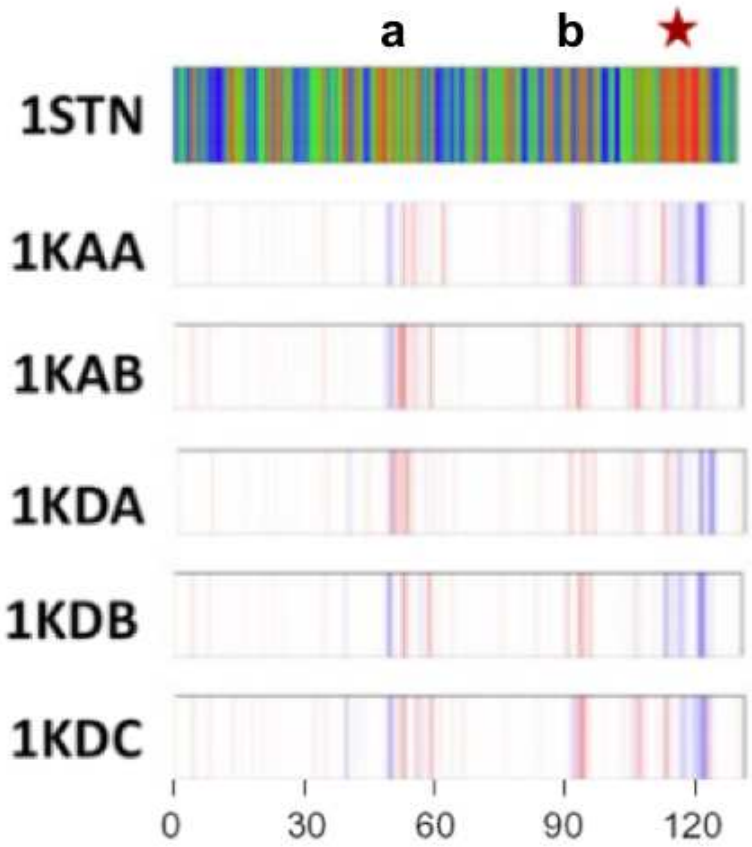

Figure 8: The color strip representation of the wild type staphylococcal nuclease is shown at the top. Single point mutation is applied on the lysine 116. 1KAA is K116A mutant; 1KAB is K116G mutant; 1KDA is K116D mutant; 1KDB is K116N mutant; and finally 1KDC is K116E mutant. Color coding for color strip is: blue, D_2 score close to 3; red, D_2 score close to -3; green, D_2 score close to 0. Color coding for color strip difference is: blue, D_2 difference value close to 6; red, D_2 difference value close to -6; green, D_2 difference value close to 0.

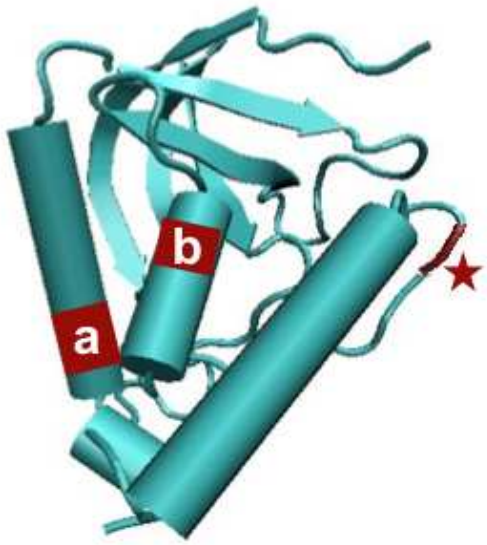

Figure 9: The tertiary structure of the staphylococcal nuclease is represented in cartoon form. Red star locates the mutation site; a is the location of 49-55 region; b is the location of 91-95 region as marked in Figure 8

2.5.4 D_2Check web server

The D_2Check function has recently been incorporated into a web-based server so as to readily allow its use in the analysis of existing PDB structures or new structures at the residue and protein scales. The server also provides related information such as the Ramachandran plot and ψ_i-ϕ_{i+1} plot of a protein structure. The values of D_2Check at the residue scale can easily be summarized using a novel color strip [84] that is also generated by the server.

All of the server-side interface and file handling software has been written in Perl, PHP and/or HTML. The user provides structural information either through a direct upload of a file that follows the PDB file protocol, or by inputing the PDB ID of the protein of interest. The default output includes the protein's Ramachandran Plot, its ψ_i-ϕ_{i+1} plot, the D_2 check value for the whole protein, and a color strip summarizing the D_2 check values at the residue scale in a compact form. Several options are available to the user: one can request analysis of a particular chain within

the structure, elect to receive the source files for the ϕ_i-ψ_i distribution, and/or elect to receive the source file for the D_2Check values at the residue scale. Some, or most, of this data is provided (once available) by way an e-mail alert message containing an HTML link to a page with a randomized address. (The results are deleted within 7 days on the server and are unlikely to be found by a webbot. Nevertheless, in future versions, these pages will use security protocols to further ensure the protection of user data from other users.)

The underlying server-side engines driven by D_2Check were developed in the Hernandez group. These codes, written in FORTRAN, perform various calculations using the structural information available in the PDB file. It is by the use of these engines that D_2Check obtains the ψ_i-ϕ_{i+1} and ϕ_i-ψ_i distributions for a given structure, references the library of distributions precomputed from the PDB, calculates a structure entropy of a given protein chain, obtains the information-theory entropies, and produces the D_1 and D_2 scores.

The D_2Check server can be used to obtain the D_2Check values at the residue scale —when analyzing the dihedral angles between and about residues— and at the protein scale —when averaging over the entire structure. In the former case, it is convenient to provide a simple and compact visualization of the $2n-3$ values. To this end, a color strip is produced by the server that uses a succession of colors to represent the values along the protein chain, starting with the N-terminus.

The server also allows the user to make requests for multiple file (or multiple chain) processing and the systematic detailed comparison of two different sequences, including a difference strip of the corresponding D_2Check values at the residue scale.

2.6 Conclusion

D_2 is a recently proposed checking function by Hernandez and coworkers [84] Although, it is not a direct scoring function for assessing structural fidelity, it does

provide insight on correlations within a chain as it combines the information from ψ_i-ϕ_{i+1} angle pairs with the information from ϕ_i-ψ_i pairs. The origin of unusual deviations signaled by the D_2 checking function [84] has been explained using the residue level analysis. The residue level analysis is complemented with the introduction of a unique colored diagram called color strip. Color strip essentially converts the values of the residual D_2 scores to visually appealing graphical representation. Although it presents individual residual D_2 scores, it provides fast and accurate interpretation to the typicality/atypicality of the overall protein structure. Because, it has been observed that a protein structure has absolute D_2 value greater than 3 only if most of the residues in its structure has absolute D_2 value greater than 3. Therefore, if color strip displays an image quite rich in red and blue (i.e. -3 and +3, respectively), it means that it is signaling for an atypical structure. Figure 7 gives several examples to typical and atypical structures in terms of color strip appearance. Color strip difference analysis, on the other hand, can only be used when comparing two structures of the same length. Exemplary tests may include characterization of native structure compare to mutants of the corresponding structure, or structure verification of computationally or experimentally generated decoys. Characterization of the relation between native structure to mutant structures, has been addressed in this section using staphylococcal nuclease and single point mutants of it at LYS116.

CHAPTER III

ADAPTIVE STEERED MOLECULAR DYNAMICS

3.1 Overview

The domain of the energy landscape of even a small peptide such as neuropeptide Y or even decaalanine has a high dimensionality. The identification of an unfolding pathway is therefore useful because it greatly reduces this dimensionality. Once identified, the energetics along this pathway is determined by the so-called potential of mean force (PMF) [133]. The importance of the PMF as well as the difficulty in calculating it has led to the development of far too many approaches to list here. Instead, we focus on those approaches which rely on sampling the states directly from trajectories. Unfortunately, the use of unconstrained trajectories is cost prohibitive when the processes of interest are very slow and dominated by deep minima. Instead, SMD can accelerate such processes by applying steering forces along the chosen unfolding pathway. Such a non-equilibrium process would not seem to provide the unconstrained structures required to obtain the equilibrium PMF. This problem was resolved by Jarzynski when he showed that an appropriately weighted average of the non-equilibrium work over many such SMD trajectories leads to the PMF [134, 135].

Jarzynski's equality has been validated numerically on several systems such as deca-alanine stretching by Park and Schulten [136], Ace-Alanine$_8$-NMe unfolding and ligand diffusion in globins by Xiong *et al* [137] and Angeli's salt decomposition by Torras *et al* [138]. It has been compared to existing biased MD techniques, such as to umbrella sampling [139] and to targeted MD [140] yielding comparable results. It has also been verified in the context of experimental results such as RNA unfolding by Liphardt *et al* [141] and a mechanical oscillator [142]. This chapter provides (i)

a review of Jarzynski's equality, (ii) the theory and a heuristic proof to the adaptive integration of Jarzynski's equality, (iii) implementation of the adaptive scheme on non-equilibrium molecular dynamics simulations.

3.2 Review of Jarzynski's equality

Jarzynski's equality was originally expressed in terms of classical Hamiltonian systems [134, 135]. It was extended to thermostated stochastic systems by Crooks [150]. Having the system in contact with a large enough heat bath such that the temperature deviation can be assumed to be nearly zero is crucial because the system will not be in equilibrium at the end of a force driven change. Crooks' introduction of a heat bath ensures that after sufficient time upon reaching a given nonequilibrium state, the system will reach an equilibrium with the environment at no additional cost of work. Jarzynski's equality for *dissipated* Hamiltonian systems can be stated as follows. Suppose a classical mechanical system consists of N particles, denoted by the phase space variables z, that are surrounded by a large enough heat bath. A constraint on the configuration space z_x is imposed through the projection $\xi_x = \xi_x(z_x)$ acting in configuration space alone. The constrained Hamiltonian may be written as:

$$H_\xi^{\mathrm{SEB}}(\Gamma, \Theta) = H^{\mathrm{SE}}(z; \Theta_x) + H^{\mathrm{B}}(\Theta) \tag{16a}$$

$$= T^{\mathrm{S}}(\xi) + H_{\xi_x}^{\mathrm{E}}(\Gamma; \Theta_x) + H^{\mathrm{B}}(\Theta) \tag{16b}$$

where S, E and B denote the constrained system, environment and bath, respectively, the subscript x (p) refers to the position (momentum) components, and T^{S} is the kinetic energy for the constrained system variables. The system variables not constrained by ξ —viz. the environment— comprise a space of dimension lower than $6N$, and its phase space variables are represented through Γ. The phase space variables Θ comprise the positions Θ_x and momenta Θ_p of the bath, and their dynamics are weakly coupled to Γ in the $H_{\xi_x}^{\mathrm{E}}$ term. The constraint ξ_x is typically one-dimensional and serves as an order parameter or reaction path that defines a state of the system.

The space defined by Γ_x is orthogonal to ξ_x and denotes the environment exclusive of the bath Θ. The non-equilibrium process between two points in the constrained space is driven by the addition of a time-dependent Hamiltonian

$$H' = H'(\xi_x, t) \tag{17}$$

that acts only on ξ_x. That is, the total time-dependent Hamiltonian is $H^{\mathrm{T}} = H_{\xi}^{\mathrm{SEB}} + H'$. In what follows, we will not generally distinguish between the phase space ξ and configuration space ξ_x variables, for simplicity.

The change in the energy as the system is carried from an initial state ξ_0 to a final state ξ_t corresponds to the work done by $H'(\xi, t)$ through this $\xi_t \leftarrow \xi_0$ process,

$$\begin{aligned} W^{\xi_t \leftarrow \xi_0}(\Gamma_t, \Theta_t, \Gamma_0, \Theta_0) &= H^E_{\xi_t}(\Gamma_t; \Theta_t) - H^E_{\xi_0}(\Gamma_0; \Theta_0) && \text{(18a)} \\ &= H'(\xi_t, t) - H'(\xi_0, 0)\ , && \text{(18b)} \end{aligned}$$

where Γ_t and Θ_t are connected to Γ_0 and Θ_0 through the propagator during the $\xi_t \leftarrow \xi_0$ process for a time t. Note that the difference in T^{S} and H^{B} vanish because of the thermalization conditions.

The equilibrium partition functions associated with the initial and final points associated with the $\xi \leftarrow 0$ process can be rewritten in terms of the original system variables as [151, 152]

$$\begin{aligned} Z_{\xi}^{\mathrm{SE}} &= \int dz e^{-\beta H^{\mathrm{SE}}(z)} \delta\left(\xi(z) - \xi\right) && \text{(19)} \\ &= \int dz d\Theta e^{-\beta\left\{H_{\xi}^{\mathrm{SEB}}(\Gamma,\Theta) + H'(\xi_x,t)\right\}} \delta\left(\xi(z) - \xi\right)\ , && \text{(20)} \end{aligned}$$

which is related to the potential of mean force, $G(\xi)$, through the reversible work theorem, $\ln Z_{\xi}^{\mathrm{S}} = -\beta G(\xi)$.

Assuming that S and B forms a canonical ensemble, the partition function of the overall system can be calculated using $Z \equiv Z^{\mathrm{S}} = \int e^{-\beta E}$. One can, therefore, relate this to the energy change defined in Equation (18) so that,

$$Z_{\xi} = Z_0 \left\langle \int e^{-\beta W_{\xi}} \right\rangle_0 \tag{21}$$

which can be further reduced to free energy representation using $G = -\frac{1}{\beta} \ln Z$

In terms of these free energies, Jarzynski's equality [134, 135, 150], is

$$G(\xi_t) = G(\xi_0) - \frac{1}{\beta} \ln \left\langle e^{-\beta W^{\xi_t \leftarrow \xi_0}} \right\rangle_0 \ , \tag{22}$$

where the ensemble average is taken over the initial variables (z, Θ) satisfying the constraint, $\xi(z_x) = \xi_0$. Note that, similar to the ground-state dominance in the calculation of a partition function, the Jarzynski average is dominated by the trajectories with the lowest work change.

Jarzynski's inequality follows from Equation (22) through the use of Jensen's inequality:

$$G(\xi_t) - G(\xi_0) \leq \left\langle W^{\xi_t \leftarrow \xi_0} \right\rangle_0 \ . \tag{23}$$

Alternatively, the use of a cumulant expression provides the second-order cumulant (SOC) expression

$$G(\xi_t) - G(\xi_0) \approx \left\langle W^{\xi_t \leftarrow \xi_0} \right\rangle_0 - \frac{1}{2}\beta \left(\left\langle \left[W^{\xi_t \leftarrow \xi_0} \right]^2 \right\rangle_0 - \left\langle W^{\xi_t \leftarrow \xi_0} \right\rangle_0^2 \right) \ , \tag{24}$$

which is surprisingly accurate for small non-equilibrium processes or environments with Gaussian response [153, 154, 136].

3.3 Adaptive scheme for the integration of Jarzynski's equality

As will be seen in Chapter IV and V, the application of the Jarzynski Equality for the extended motion of a finite number of unfolding trajectories provides a very weak upper bound to the PMF. In fact, it is so weak that the cumulant expansion of Equation (22) presents a dramatically large deviation between the second order cumulant and the exponential average as demonstrated in Figure 24. The PMF can be converged by several plausible but highly expensive ways including significantly increasing the sample size, decreasing the pulling velocity, and equilibrating the system

at short intervals. Instead, in order to treat such extended systems without increasing the computation cost, we have developed anadaptive version of Schultens' algorithm [136] in which the Jarzynski equality is applied through a series of shorter steps. It is *adaptive* in the sense that the initial configuration for a given step is obtained (or adapted) from the trajectories of the previous step.

The overall unfolding path is initially partitioned into N steps marked by its endpoints, $\xi_0, \xi_1, ..., \xi_N$. The i^{th} iteration is initiated at ξ_{i-1} and Γ_{i-1} while the bath Θ_{i-1} is sampled from the appropriate canonical ensemble. Each such bath, $\Theta_\alpha^{\xi_i \leftarrow \xi_{i-1}}(t_{i-1})$, leads to M trajectories labeled by α for the $\xi_i \leftarrow \xi_{i-1}$ process. This in turn, leads to a distribution of values in the work $W_\alpha^{\xi_i \leftarrow \xi_{i-1}}(t)$, environment $\Gamma_\alpha^{\xi_i \leftarrow \xi_{i-1}}(t)$, and bath $\Theta_\alpha^{\xi_i \leftarrow \xi_{i-1}}(t)$ for times t within the i^{th} step. At the end of the iteration, the average work $W^{\xi_i \leftarrow \xi_{i-1}}(t_i)$ is computed according to the Jarzynski equality (Equation (22)). There then exists a trajectory α' for which its work $W_{\alpha'}^{\xi_i \leftarrow \xi_{i-1}}(t_i)$ is closest to the average work $W^{\xi_i \leftarrow \xi_{i-1}}(t_i)$. The initial value of the environment Γ_i for the $(i+i)^{\text{th}}$ iteration is then taken to be the corresponding $\Gamma_{\alpha'}^{\xi_i \leftarrow \xi_{i-1}}(t_i)$. Meanwhile the algorithm is initiated with values (ξ_0, Γ_0) matching the initial structure of the system and environment. This amounts to the structure of the entire protein while ξ refers only to the constrained angle spanned by the helix and tail.

A proof of this algorithm begins by considering the application of the adaptive procedure to divide a single step into two substeps as illustrated in Figure 10 along a specific unfolding path λ for the corresponding system variables ξ. For simplicity, but without loss of generality, we suppose that the the system is carried along by a nonequilibrium process from state $\xi = 0$ at initial time 0 to a final state $\xi = 1$ at a final time t. For each of M realizations labeled by α of the $1 \leftarrow 0$ process, the trajectories of the environment $\Gamma_\alpha^{1\leftarrow 0}(t)$ and the bath $\Theta_\alpha^{1\leftarrow 0}(t)$ can be formally constructed. The work done along each of these trajectories is $W_\alpha^{1\leftarrow 0}[\Gamma_\alpha^{1\leftarrow 0}(t), \Theta_\alpha^{1\leftarrow 0}(t), \Gamma_\alpha^{1\leftarrow 0}(0), \Theta_\alpha^{1\leftarrow 0}(0)]$ as

specified by Equation (18). The PMF of this process is

$$\Delta G^{1\leftarrow 0} = -\frac{1}{\beta} \ln \left\langle e^{-\beta W_\alpha^{1\leftarrow 0}} \right\rangle_\alpha \tag{25}$$

where the average is taken over the M realizations starting with the same initial ξ_0 and Γ_0 and various initial bath configurations $\Theta_\alpha(0^{1\leftarrow 0})$.

The single step can now be partitioned into two steps in which the system is stopped at an intermediate time t' and the corresponding position ξ'. For each of the original M trajectories in the $1 \leftarrow 0$ process, this partitions the work into two components:

$$W_\alpha^{\xi'\leftarrow 0} = H_{\xi'}^{SB}[\Gamma_\alpha^{1\leftarrow 0}(t'), \Theta_\alpha^{1\leftarrow 0}(t')] - H_{\xi'}^{SB}[\Gamma_\alpha^{1\leftarrow 0}(0), \Theta_\alpha^{1\leftarrow 0}(0)] \tag{26a}$$

$$W_\alpha^{1\leftarrow \xi'} = H_{\xi'}^{SB}[\Gamma_\alpha^{1\leftarrow 0}(t), \Theta_\alpha^{1\leftarrow 0}(t)] - H_{\xi'}^{SB}[\Gamma_\alpha^{1\leftarrow 0}(t'), \Theta_\alpha^{1\leftarrow 0}(t')] \tag{26b}$$

from which the free energy change for the first step can be easily obtained using Jarzynski's equality

$$\Delta G_{\xi'\leftarrow 0} = -\frac{1}{\beta} \ln \left\langle e^{-\beta W_{\xi'\leftarrow 0}^{\alpha}} \right\rangle_\alpha \ . \tag{27}$$

For the second substep, however, each trajectory specified by Equation (26b) starts at a different value of the environment, $\Gamma_\alpha^{1\leftarrow 0}(t')$. We now introduce a $\xi' \leftarrow \xi'$ process during which ξ' is held fixed and the environment $\Theta_\alpha^\tau(t')$ relaxes in time τ from 0 to τ_α for some arbitrary final time τ_α which is likely different for each trajectory α. The work to move the system from the state at the end of the process described in Equation (26a) along this $\xi' \leftarrow \xi'$ process is

$$\Delta W_\alpha^{\xi'\leftarrow \xi'} = H_{\xi'}^{SB}[\Gamma_\alpha^{\tau_\alpha}(t'), \Theta_\alpha^{\tau_\alpha}(t')] - H_{\xi'}^{SB}[\Gamma_\alpha^{\xi'\leftarrow 0}(t'), \Theta_\alpha^{\xi'\leftarrow 0}(t')] \ , \tag{28}$$

and the work to return to the final point of the $1 \leftarrow 0$ process is

$$W_\alpha^{'\,1\leftarrow \xi'} = H_{\xi'}^{SB}[\Gamma_\alpha^{1\leftarrow \xi'}(t), \Theta_\alpha^{1\leftarrow \xi'}(t)] - H_{\xi'}^{SB}[\Gamma_\alpha^{\tau_\alpha}(t'), \Theta_\alpha^{\tau_\alpha}(t')] \ . \tag{29}$$

The $\xi' \leftarrow \xi'$ process can be allowed to propagate for as long as it takes for $\Gamma_\alpha^{\tau_\alpha}(t')$ to be equal to some $\Gamma^{1\leftarrow \xi'}(t')$ which is independent of α. The existence of such a

common endpoint is assured if the process is ergodic and the system is found in a single local basin of attraction. The requirement of ergodicity is a weak constraint given that the environment is coupled to a bath. The requirement for a single basin is also weak because the environment must access all possible such basins with zero-work paths. This motivates a new path for a *restricted* $1 \leftarrow \xi'$ process starting at the fixed endpoint $\Gamma^{1\leftarrow\xi'}(t')$, and its work is given by

$$W_\alpha^{\prime\prime\, 1\leftarrow\xi'} = H_{\xi'}^{SB}[\Gamma_\alpha^{1\leftarrow\xi'}(t), \Theta_\alpha^{1\leftarrow\xi'}(t)] - H_{\xi'}^{SB}[\Gamma^{1\leftarrow\xi'}(t'), \Theta_\alpha^{1\leftarrow\xi'}(t')] \,, \tag{30}$$

where the stochastic $\Theta_\alpha^{1\leftarrow\xi'}(t')$ has replaced the formally propagated $\Theta_\alpha^{\tau}(t')$. That is, the bath decoherence time is sufficiently fast that the detailed propagation can be ignored while the initial bath $\Theta_\alpha^{1\leftarrow\xi'}(t')$ in the $1 \leftarrow \xi'$ process is Gaussian random. The PMF of the restricted $1 \leftarrow \xi'$ process is

$$\Delta G^{1\leftarrow\xi'} = -\frac{1}{\beta} \ln \left\langle e^{-\beta W_\alpha^{\prime\prime\, 1\leftarrow\xi'}} \right\rangle_\alpha \tag{31}$$

The average in Equation (25) can thus be written as:

$$\left\langle e^{-\beta W_\alpha^{1\leftarrow 0}} \right\rangle_\alpha = \left\langle e^{-\beta\left\{W_\alpha^{\prime\prime\, 1\leftarrow\xi'} + \Delta W_\alpha^{\xi'\leftarrow\xi'} + W_\alpha^{\xi'\leftarrow 0}\right\}} \right\rangle_\alpha \,, \tag{32}$$

where it should be noted that the sum in the exponent in the RHS is not equal to $W_\alpha^{1\leftarrow 0}(t)$ nor is the trajectory the same after t'. However, the averages are equal because they are both non-equilibrium $1 \leftarrow 0$ processes between the same initial and final points satisfying Jarzynski's equality. Meanwhile the work in the $\xi' \leftarrow \xi'$ process is zero because the system was allowed to relax freely. Hence,

$$\left\langle e^{-\beta W_\alpha^{1\leftarrow 0}} \right\rangle_\alpha = \left\langle e^{-\beta\left\{W_\alpha^{\prime\prime\, 1\leftarrow\xi'} + W_\alpha^{\xi'\leftarrow 0}\right\}} \right\rangle_\alpha \tag{33a}$$

$$= \left\langle e^{-\beta W_\alpha^{\prime\prime\, 1\leftarrow\xi'}} \right\rangle_\alpha \times \left\langle e^{-\beta W_\alpha^{\xi'\leftarrow 0}} \right\rangle_\alpha \,. \tag{33b}$$

The second equality follows from the fact that the trajectories in the $1 \leftarrow \xi'$ and $\xi' \leftarrow 0$ processes are uncoupled and independently sampled. Combining Equations (25),

(27), (31) and (33b), we obtain the desired result,

$$\Delta G^{1\leftarrow 0} = \Delta G^{1\leftarrow \xi'} + \Delta G^{\xi'\leftarrow 0} \,, \tag{34}$$

where the initial value of the environment $\Gamma^{1\leftarrow\xi'}(t')$ at the beginning of the $1 \leftarrow \xi'$ process, in principle, can be chosen to be any arbitrary (but the same) state that is accessible to a $\xi' \leftarrow \xi'$ process. However, the choice of that intermediate state will affect the accuracy and convergence of the approach in so far as better choices would be more easily accessible and thus require less numerical relaxation in the evolution of the $1 \leftarrow \xi'$ process. The best such choice is one that corresponds to a typical structure (not the minimum energy state) associated with the non-equilibrium process. To this end, we choose $\Gamma^{1\leftarrow\xi'}(t')$ according to the $\Gamma_\alpha^{1\leftarrow\xi'}(t')$ corresponding to the trajectory α which minimizes the work difference, $\left|\Delta G^{\xi'\leftarrow 0} - W_\alpha^{\xi'\leftarrow 0}(t')\right|$.

Repeated application of Equation (34) and the associated proscription for the choice of intermediate environment variables Γ for N steps gives rise to the desired final expression for the adaptive free energy difference:

$$\Delta G = \sum_{i=1}^{N} \Delta G^{i\leftarrow(i-1)} \,, \tag{35}$$

where i labels the corresponding steps. In the limit that the "environment variables" are empty—i.e., that the dimensionality of the Γ space is zero— the adaptive procedure reduces to the use of the Jarzynski equality with the additivity trivially arising from the fact that the free energy is a state function.

In so far as the bath has been assumed to be Gaussian, the adaptive procedure should fail if the second-order cumulants in the work of a given set of trajectories begin to be nontrivial. As is shown below, the adaptive procedure does indeed satisfy this requirement.

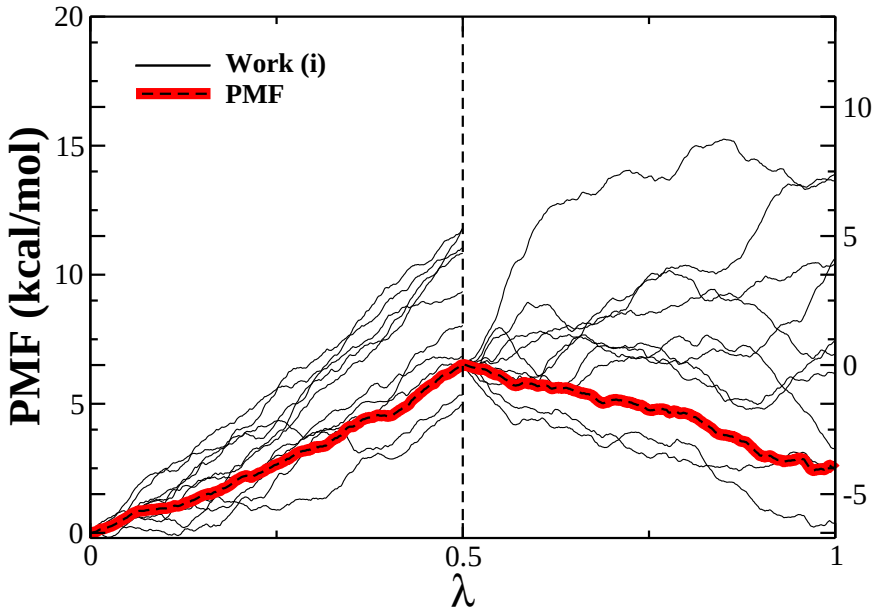

Figure 10: Illustration of the adaptive scheme applied to a system where the unfolding path is divided into two steps (with 10 trajectories α each): Black solid curves are the work for each of the 10 trajectories at each step. The PMF along a given substep is shown with a thick-red highlighted black-dashed cure. The right y-axis tick marks are labeled for the second-step work trajectories with the 0 position located at the final average work value of the first substep. The left y-axis tick marks are labeled for both the first step work trajectories and the overall PMF. (This figure is drawn for illustration purpose only and is not based on real physical data.)

3.4 Adaptive steered molecular dynamics

The external forces that carry the system along a particular reaction coordinate, $\xi(z_x)$, are imposed by way of a predefined potential $H'(\xi(z_x);\lambda)$. With the addition of this new potential, the extended time-dependent Hamiltonian, H^{ext}, becomes:

$$H^{\text{ext}}(z,\Theta;\lambda) = H_{\xi}^{\text{SEB}}(\Gamma,\Theta) + H'(\xi(z_x);\lambda) \tag{36}$$

where

$$H'(\xi(z_x);\lambda) = \frac{1}{2}k[\xi_x(z_x) - \lambda]^2 \ , \tag{37}$$

for a specified time-dependent process $\lambda(t)$. In a typical steered molecular dynamics simulation pulling is achieved by harmonically (Equation (37)) attaching a specific atom or group of atoms to an auxiliary particle and steering this imaginary particle along the desired path. In the adaptive scheme, the pulling process is staged in N linear steps so as to approximate the complete the whole reaction coordinate. Thus for each step i, the auxiliary potential in Equation (37) becomes

$$U_i(\vec{r}) \ = \ \frac{1}{2}k\left[\vec{r}(t) - (\vec{r}_i + v_i\vec{n}_i t)\right]^2 \ , \tag{38}$$

where $\vec{r}_i$ is the position of the center of mass of the steered atom(s) at the beginning of the interval, v_i is the velocity to move the particle to the end in the fixed time step, and $\vec{n}_i$ is the direction between the initial and final positions of the steered atom(s). The position $\lambda_i \equiv (\vec{r}_i + v_i\vec{n}_i t)$ can be associated with the pseudo particle (or dummy atom) that follows smoothly the prescribed unfolding path. As it does so, it exerts a work on the system ξ that is given by

$$\Delta W_i(t) = \int_{t_{i-1}}^{t} \vec{F}_i \cdot \vec{n}_i v dt \ , \tag{39}$$

where the force $\vec{F}_i = -\nabla U_i(\xi_x(z_x))$ is related to the corresponding potential of Equation (38). The corresponding free energy change, $\Delta G^{t\leftarrow 0}$, at time t within the i^{th}

interval can now be calculated using the adaptive work expression in Equation (35), *i.e.*,

$$e^{-\beta\Delta G_{t\leftarrow 0}} = \left\langle e^{-\beta\Delta W_i(t)} \right\rangle_i \times \prod_{j=1}^{i-1} \left\langle e^{-\beta\Delta W_j(t_j)} \right\rangle_j \ , \tag{40}$$

where the subscript on the angle brackets denotes the averaging over the trajectories in the corresponding interval.

Formally, the initial configuration for each subsequent iteration i can be obtained by holding the perturbation fixed—that is λ held at ξ_i—long enough that the environment relaxes to equilibrium. During this waiting period, no work is done on the system, and hence there is no contribution to Equation (39). Thus Equation (40) is a formally exact way of restating Jarzynski's equality in a series of steps. In such an implementation it offers little computational advantage in so far as the relaxation stages could be quite expensive. It does, however, offer an advantage in the convergence as the trajectories are less likely—because they are shorter—to wander off to distant parts of the landscape. The statistics of the work distribution is consequently more nearly Gaussian, and the convergence of the sum is faster. Echevarria and Amzel [155] essentially followed this procedure in obtaining the helix propensities of dodecaalanine in solvent using a small number (15) of trajectories allong a 15Å stretch.

The computational advantage is potentially greater, however, if a more computationally efficient criteria can be applied to the choice of the starting configuration at each step. Possible choices are the configuration that requires the minimum amount of work, that is nearest to the reaction coordinate at the end of the iteration, or that requires the amount of work that is closest to the Jarzynski's average. The last of these was the choice that was used in Ref. [143] and it is confirmed as the best choice among these in the case of the stretching of decaalanine as displayed in Section 4.3.1. Intuitively, it makes the most sense because it amounts to selecting a structure that is fully relaxed—taking advantage of the results from the trajectories that have already

been calculated—without requiring an additional relaxation period before initiating the next batch of trajectories.

3.5 Conclusion

Calculating the free energy profile of biophysical experiments has been a major challenge since it requires the sampling of the whole configuration space, which is extremely costly in both numerical and experimental studies. One way to reduce the cost is to superimpose a time-dependent force such as in umbrella sampling, free energy perturbation theory, weighted histogram analysis method (numerical), optical tweezers, atomic force microscopy (experimental). After Jarzynski introduced the non-equilibrium work relation (i.e. Jarzynskis equality) in 1997 [134], steered molecular dynamics (SMD) has been shown to efficiently calculate the free energy of processes along the steering path [136]. Adaptive steered molecular dynamics [143] is a staged algorithm, which works by simply dividing the reaction coordinate of interest in multiple steps and —at each step— sampling the configuration space over a selected final configuration that is obtained at the end of the preceding step. As will be seen in Chapters IV and V, the adaptive scheme is found not only to improve the efficiency even further but also to provide a more accurate free energy profile with narrower error bars.

CHAPTER IV

THE ENERGETICS OF DECA-ALANINE STRETCHING IN WATER OBTAINED BY ADAPTIVE STEERED MOLECULAR DYNAMICS SIMULATIONS

4.1 Overview

Investigating biological reactions *in silico* is a major scientific challenge since such processes generally involve large conformational changes in biomolecules (proteins, nucleic acids etc.) and occur in time scales of milliseconds to tens of seconds. Unfortunately, current simulation methods and hardware capacities allow modeling of such processes for up to hundreds of nanoseconds. Although lacking in the same context, molecular dynamics (MD) is generally accepted to yield accurate results. Thus, many research is dedicated towards developing strategies within MD to improve the efficiency of simulations including but not limited to replica exchange MD [128], adaptive biasing force MD [131], free energy perturbation MD [157, 158]. Amongst the biased integration methods, steered molecular dynamics (SMD) in corporation with the Jarzynski's non-equilibrium work relation [134, 135] has been shown to accurately predict free energy profile of bioprocesses along a predefined steering path (i.e. the reaction coordinate).

Providing an exact relation between the free energy difference and the work done through a directed non-equilibrium process, Jarzynski's equality has been validated in the context of both experimental studies [141, 142] and computational reports [150, 159, 160, 161, 162, 163]. The aim of the study presented herein is twofold: (i) reproduce the free energy profile of deca-alanine stretching in vacuum [136] using adaptive steered molecular dynamics [143] with significantly less CPU time utilized

(ii) extend the investigation to in solvent stretching of deca-alanine to explore the energetics of the hydration effect.

4.2 Model and methods

Studies [164, 165, 166, 167, 168, 169] of small molecules—e.g., peptides—continue to play a key role as benchmarks of new methods and to test fundamental hypotheses in the context of protein dynamics and folding. While necessarily small, the deca-alanine peptide contains several internal hydrogen bonds that must be broken in order to fully stretch it. As such, it is a simple target for demonstration and verification of methods that probe the energetics of unraveling a peptide. It is for this reason that the seminal work on the unraveling of deca-alanine in vacuum [136, 170] was so instructive, and why doing so in a water solvent provides a large test to emerging methods.

Molecular dynamics simulations have been carried out using NAMD [171] with the CHARMM force field [172]. Decaalanine is fully specified in terms of an 104-atom model with the hydrogens included explicitly. Water molecules have been treated using the TIP3P [173] potential model within NAMD. The use of more sophisticated water models may affect the results but they would also require more computational resources. Simulations of the peptide in solvent were carried out at 300 K in a box of dimensions 52Å $\times$ 52Å $\times$ 65Å filled with 5,138 water molecules. The peptide is stretched along the longest of these lengths, and thereby ensures that it does not see itself or water layers affected by its motion. Temperature is controlled using Langevin dynamics as implemented in the built-in NAMD integrator.

The simulation box is initially thermally equilibrated by first heating the bath from 0 K to 300 K in 5 ps. The appropriate density at the given simulation conditions is subsequently obtained using a constant pressure (NPT) propagation for 5 ps. An additional equilibration at constant volume and temperature is then performed for

10 ps.

4.2.1 Adaptive steered molecular dynamics of the deca-alanine stretching

The unraveling of deca-alanine is investigated using the adaptive SMD algorithm as the peptide is pulled apart from the ends at different velocities (i) in vacuum so as to compare the standard SMD simulations of Schulten *et al*, and (ii) in solvent so as to observe the effect of hydration on helix-coil transition.

The reaction coordinate is defined as the end-to-end distance between the nitrogen atom of the N-terminus (NN) and the nitrogen atom of the cap at the C-terminus (NC). The pulling of the peptide is imposed using the steering module within NAMD by holding the NN end fixed and directing the NC end relative to the NN end. The overall unraveling coordinate covers the NN-NC distance from 13Å to 33Å (Figure 11). Adaptive steered molecular dynamics for the stretching of deca-alanine is designed to cover this coordinate in 10 steps. In order to accurately measure the PMFs obtained from adaptive SMD simulations, standard SMD simulations in vacuum are also implemented.

The system is simulated at various pulling velocities in vacuum and in solvent. Trajectories are analyzed both numerically and empirically. The latter is facilitated by the NAMD/VMD package. PMF's along the deca-alanine stretching pathway are calculated for each set of simulations. The effect of the solvent is also investigated by the hydrogen bond count within the α-helix and between the α-helix and water molecules around it. These are identified according to a simplified criteria: Hydrogen bonds are identified when one of two nonbonded pair of atoms at $\vec{r}_1$ and $\vec{r}_3$ is bonded to a hydrogen atom at $\vec{r}_2$, the distance $|\vec{r}_3 - \vec{r}_1|$ is less than 4.2Å and the angle spanned by the rays from $\vec{r}_2$ to the nonbonded atoms is less than 40°.

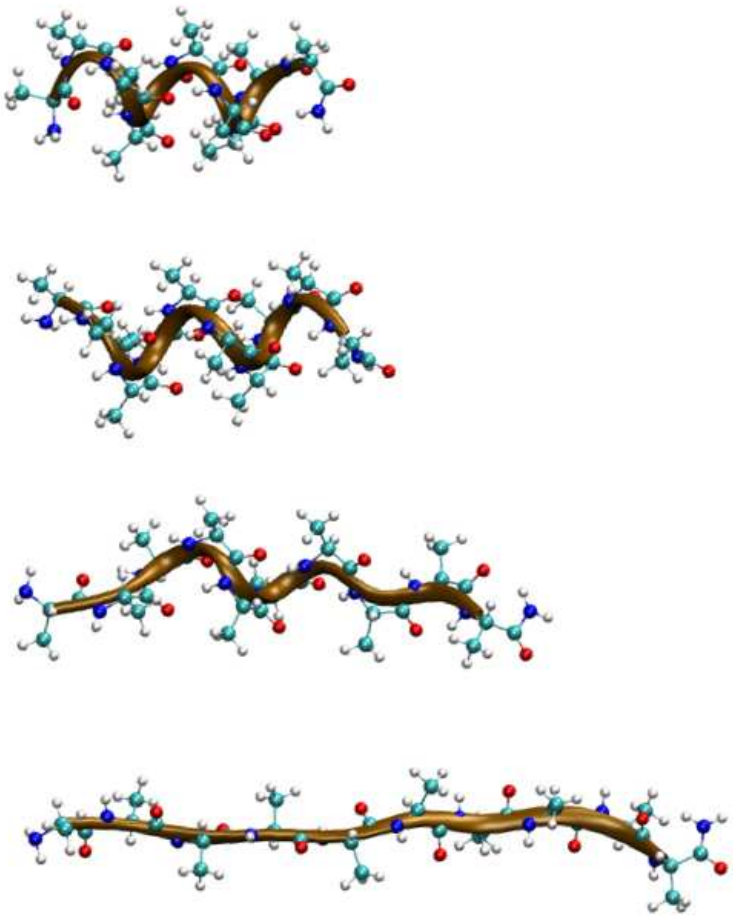

Figure 11: Ribbon and atomically detailed snapshots of deca-alanine as it is pulled in vacuum along its ends. The top image is a representative equilibrium compact structure; the NN-NC distance is 13Å. The bottom image shows the coil structure at the end of one of the pulled trajectories an end-to-end distance of 33Å. The second image from the top is the minimum energy conformation —an α-helix with an end-to-end distance of 15.2Å. The third from the top shows a representative structure and the end-to-end distance —circa 26Å–of the kink seen in the PMF shown in Figure 14.

4.3 Results and discussion

4.3.1 On the criteria for selecting the initial adaptive SMD configuration at each iteration

When choosing the configuration as the input for each stage of an adaptive steered molecular dynamics simulation one could in principle define several criteria for comparing the final configurations obtained at each step. In principle, any such structure will relax to match the correct trajectories during the nonequilibrium stretch. The average of these structures will lead to a converged result given sufficient averaging, and assuming that the initial configurations are not biased in some way. One way to generate the structure is thus to simply allow all the trajectories to relax for a finite time during which no work is done on the system. However this is potentially cost-prohibitive. Thus the determination of a suitable criteria for choosing a structure—assuming that it is not biased—provides a significant possible savings in computational effort.

The three most promising criteria that have been explored are:

(i) The configuration that requires the amount of work closest to Jarzynski's average (JA).

(ii) The configuration that requires the lowest amount of work (MW).

(ii) The configuration that is the nearest to the reaction coordinate (RC).

The PMFs in Figures 12 and 13 are calculated using a pulling velocity of 100Å/ns and 10Å/ns, respectively. In both figures, the top, middle, and bottom panels correspond to the JA, MS and RC criteria, respectively.

The choice of MW tends to distort the average towards lower values as the number of trajectories increases for a fixed pulling velocity. This is evident in Figures 12 and 13 with the increasing number of tps leading to a worse result rather than convergence to the correct result. In principle, a slower pulling velocity would narrow

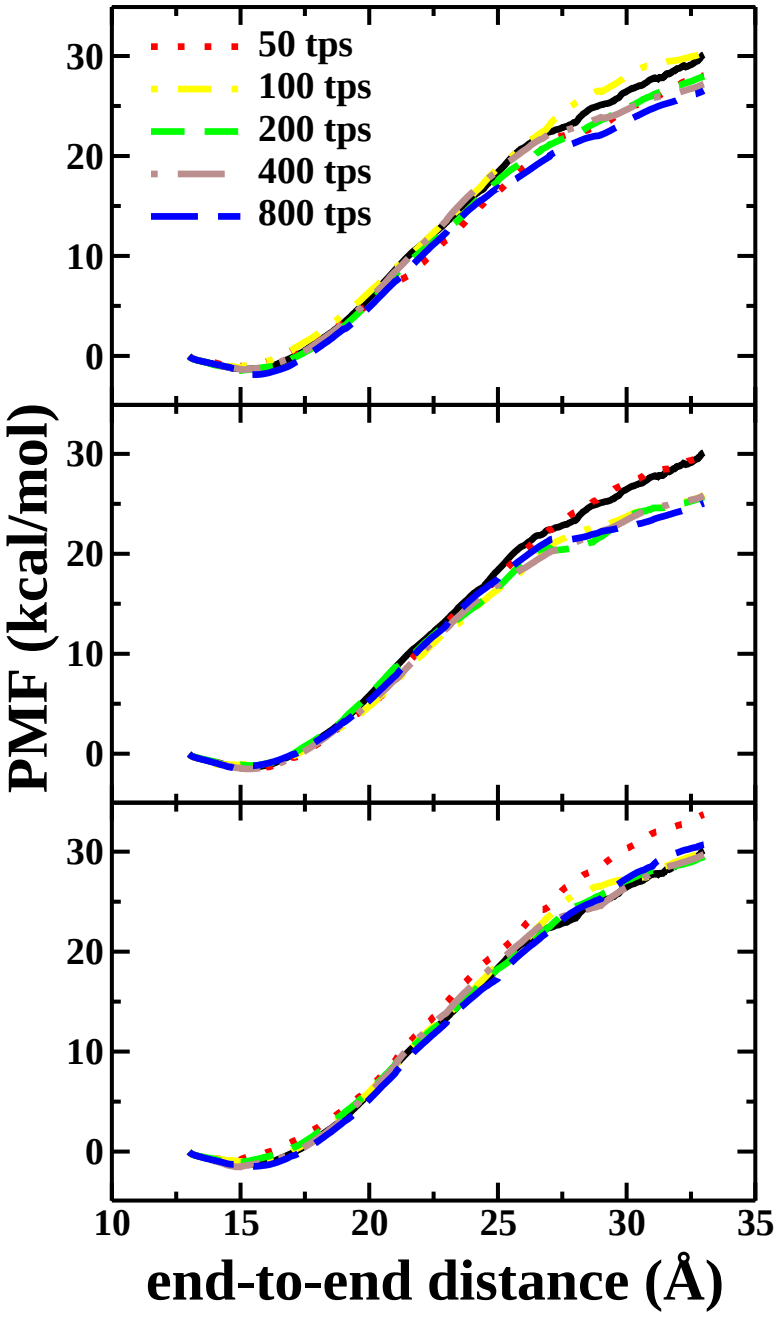

Figure 12: The comparison of the PMFs obtained from the adaptive SMD method pulling at 100 Å/ns when different selection criteria are used to choose the configuration from the structures at the end of each step. The configuration is chosen according to one of the following criteria: (i) the one that requires the amount of work closest to Jarzynski's average (JA) (bottom panel), (ii) the one that requires the lowest amount of work (MW) (middle panel), and (iii) the one that is nearest to the reaction coordinate (RC) (top panel). The legend defines the labeling of the curves according to the number of trajectories per step (tps). The solid black curve is the exact PMF obtained from averaging 10,000 standard SMD simulations.

the trajectories such that the minimum trajectory would remain within a fluctuation of the reversible ensemble. Consequently, this should formally lead to a reasonable result given sufficiently slow pulling. However, the numerical results clearly indicate that this would be cost prohibitive.

The choice of the RC leads to averages that oscillate unpredictably around the averaged work. This choice is tantamount to moving the ensemble positions back to pulling path. As such, it is effectively doing work to move the position but the work associated with this move is not accounted for. As this work can be positive or negative, it leads to the seemingly random increases and decreases in the PMF. The accuracy should improve, however, with increasing TPS and slower velocity. As the particle is forcibly stretched at slower velocities, the swarm of trajectories will track the path better, and the magnitude of the unaccounted work will decrease. However, like the MW choice, this convergence will be slow and cost prohibitive.

Finally, the choice of JA leads to averages that clearly converge well to the SMD results for a given pulling velocity, and in particular to the exact reversible work when the SMD is slow enough. In these cases, it occurs with as little as 400-800 trajectories per step which is substantially less than the 10,000 trajectories that are required to converge the standard SMD. Thus the adaptive SMD using the JA criterion appears to be both accurate and efficient in terms of CPU requirements. The adaptive SMD simulations that are presented through the remainder of this thesis are realized by utilizing the choice of JA.

4.3.2 Helix-coil transition of the deca-alanine in vacuum

The forced unraveling of deca-alanine is first investigated in vacuum at two pulling velocities (10Å/ns and 100Å/ns). The adaptive SMD simulations are performed in 10 incremental steps as this was found to be sufficient to obtain convergence. This covers a change in the overall stretching coordinate of 20Å—that is, the NN-NC distance

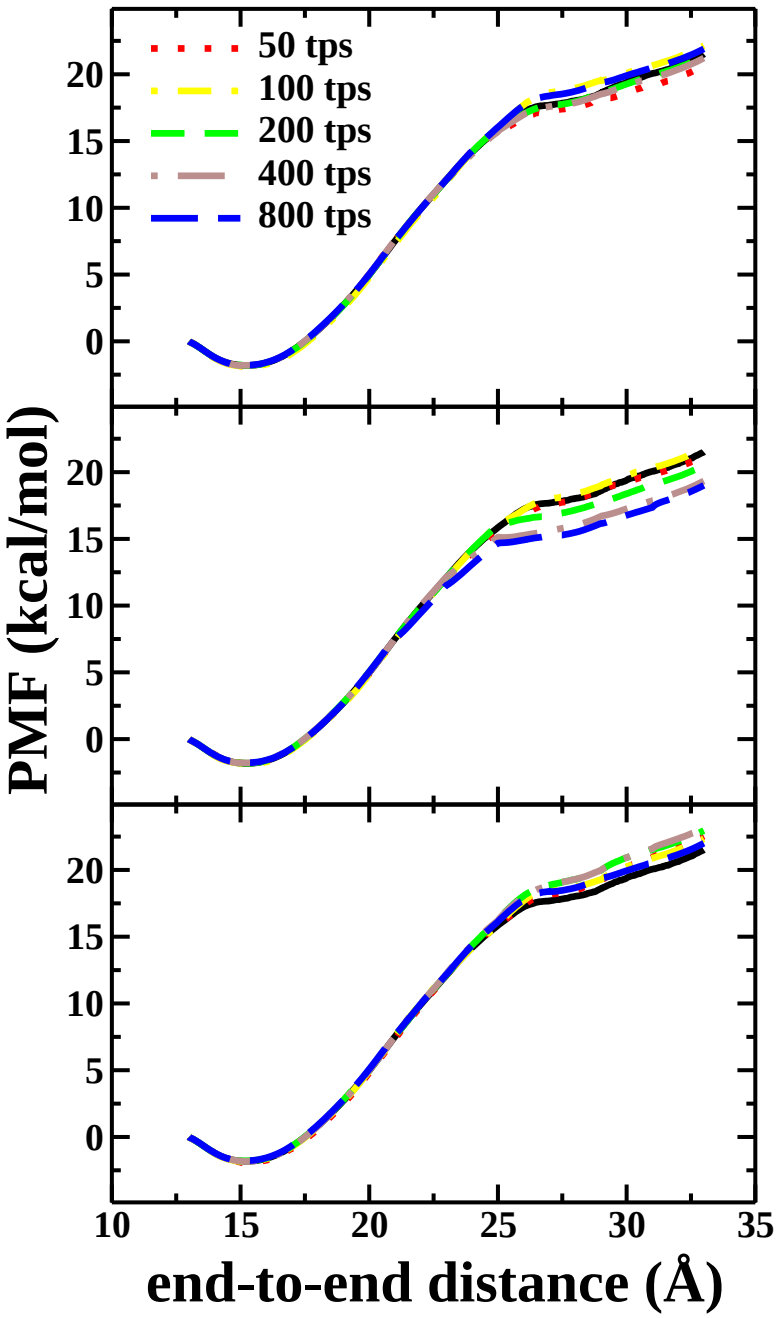

Figure 13: The comparison of the PMFs obtained from the adaptive SMD method pulling at 10 Å/ns when different selection criteria are used to choose the configuration from the structures at the end of each step. The configuration is chosen according to one of the following criteria: (i) the one that requires the amount of work closest to Jarzynski's average (JA) (bottom panel), (ii) the one that requires the lowest amount of work (MW) (middle panel), and (iii) the one that is nearest to the reaction coordinate (RC) (top panel). The legend defines the labeling of the curves according to the number of trajectories per step (tps). The solid black curve is the exact PMF obtained from averaging eight reversible (i.e. pulling velocity is 0.1 Å/ns) SMD simulations.)

goes from 13Å $\rightarrow$ 33Å). For each of pulling velocity, sets of 50, 100, 200, 400 and 800 trajectories have been simulated at each step to establish convergence with the number of trajectories. (Results using 100 and 400 trajectories per step are not shown in Figure 14 but consistent with the trends—cf. Figures 12 and 13) Note that the increase of the number of trajectories requires a nearly complete recalculation of the entire adaptive SMD. This is necessary because the configuration chosen at the end of a given iteration segment and utilized to initialize the next iteration segment is invariably different given the introduction of more trajectories. For exact comparison we have also reproduced the PMF obtained by Park and Schulten [136]. The black curve in Figure 14 has been obtained using 10,000 standard SMD trajectories at a forced velocity that is slow enough to be nearly reversible. It is is the same as that reported earlier within the standard error of the calculations. By definition of the Jarzynski's inequality, as the number of the irreversible trajectories increases the estimated PMF should converge towards the exact PMF obtained from reversible simulations. The catch is that the standard implementation requires many more such trajectories as the end-to-end distance is pulled farther from the original structure. As illustrated in Figure 14 the adaptive SMD algorithm, however, a relatively small number of trajectories—800–is sufficient to reproduced the PMF for the stretching of deca-alanine from a helix to a coil in vacuum.

4.3.3 Helix-coil transition of the deca-alanine in solvent

The forced unraveling of deca-alanine in solvent has also been investigated at three pulling velocities (10Å/ns, 33Å/ns and 100Å/ns). For each pulling velocity, 50, 100, 200 and 400 trajectories have been simulated at each step. The addition of 4,078 water molecules into the system box, dramatically increases the CPU time per trajectory from 200 minutes to 65 hours (when pulling at the slowest velocity, 10Å/ns). Assuming constant access to 48 standard cores, it would take over 564 days to acquire

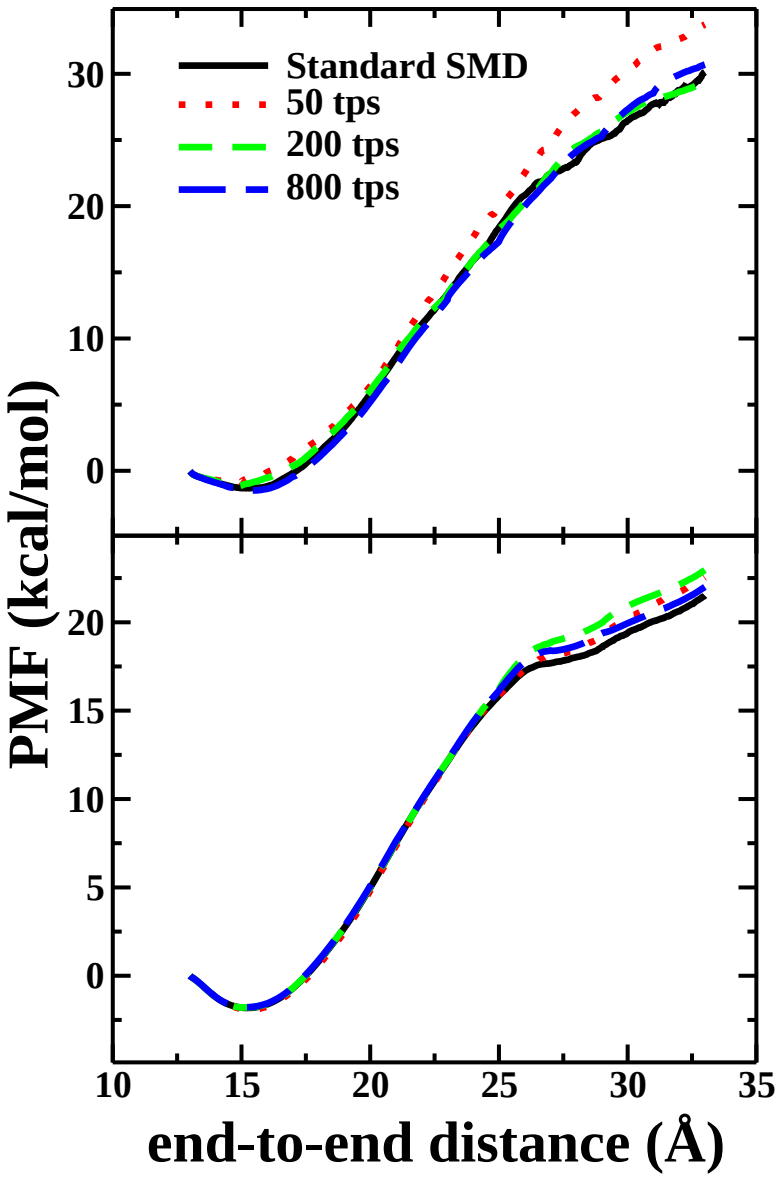

Figure 14: The comparison of the PMF's obtained from the adaptive SMD method to the PMF obtained from the standard SMD method is displayed for both pulling velocities, 100Å/ns (top panel) and 10Å/ns (bottom panel). The reversible PMF is shown as a black curve. The series of PMFs obtained using the adaptive SMD method with an increasing number of trajectories per step (tps) is labeled according to the legend.

10,000 trajectories in solvent. For the adaptive SMD calculations, it took less than 40 days to simulate all four sets of simulations (50, 100, 200 and 400 trajectories per step) at the slowest pulling velocity (i.e. 10Å/ns) using 96 cores.

The calculated PMFs for the solvated deca-alanine stretch are shown in Figure 15. The best calculated value—shown as the blue dot-dashed curve in the bottom panel—is somewhat lower in energy than the the vacuum result—shown as a solid black curve. Such a lowering is to be expected because of the stabilization provided by the water molecules as the peptide is unraveled. Meanwhile, the energetics remain structured in the sense that there is an initial fast rise—possibly two—and a subsequent slower rise after 25Å. The near agreement between the PMFs found in the initial stages of the pulling should not be surprising because they are near the original configuration and hence the underlying Gaussian behavior in the work distribution is followed well. At longer pulling distances, however, the calculated PMFs do not converge well with increasing number of trajectories per steps at the fast pulling velocities—100Å/ns in the top panel and 33Å/ns middle panel. At the lowest pulling velocity—10Å/ns in the bottom panel—the structural elements of the PMFs are independent of the sampling size and show little deviation in the free energies.

The converged PMF shown in the bottom panel of Figure 15 reveals important structural properties of the helix-coil transition of deca-alanine in water. The initial structure is evidently not the minimum energy structure as the energy minimum appears at around 14.9Å in nearly all the simulations roughly independent of pulling velocity. A small shift in the slope of the rise in the PMF appears near 20Å which is not seen in the vacuum limit. A more pronounced flattening in the PMF is seen near 25Å both in vacuum and in solvent. It is notable that the forced stretch does not exhibit these features for the high velocity cases. We hypothesize that this is due to insufficient relaxation of the water molecules around the non-equilibrium deca-alanine structure in the fast forced stretching cases. This hypothesis is corroborated by the

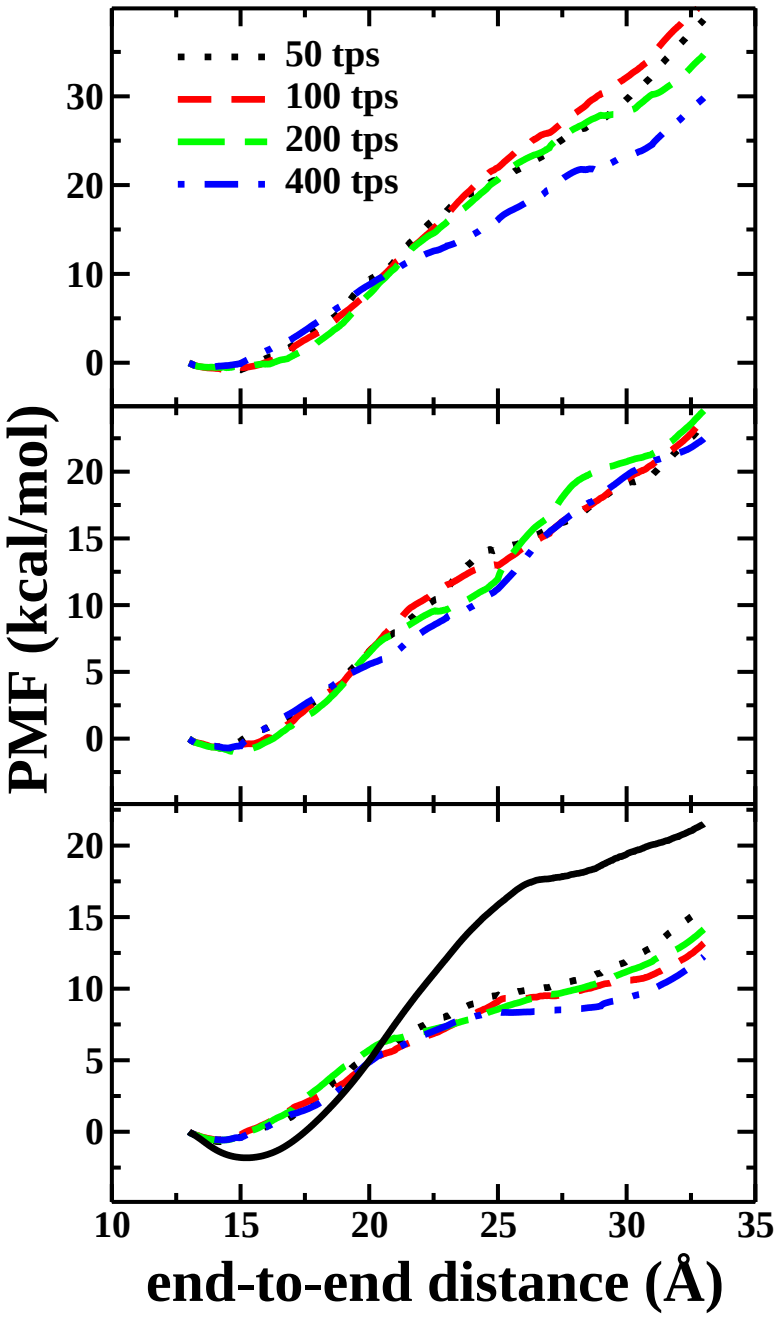

Figure 15: The PMFs obtained for the forced stretching of deca-alanine in solvent using the adaptive SMD method are displayed for different pulling velocities: 100Å/ns (top panel), 33Å/ns (middle panel) and 10Å/ns (bottom panel). The series of PMFs obtained using the adaptive SMD method with an increasing number of trajectories per step (tps) is labeled according to the legend. The solid black curve in the bottom panel reproduces the PMF obtained from the 10Å/ns adaptive SMD simulations in vacuum as shown in Figure 14.

hydrogen bond counts displayed in Figure 16. The average number of intrapeptide hydrogen bonds in vacuum and solvent are shown in the top and middle panels, respectively. The average number of interpeptide hydrogen bonds to the solvent are shown in the bottom panel. When deca-alanine unravels into a solvated coil (with no intrapeptide hydrogen bonds), the total number of hydrogen bonds between it and water is 27. This is the same as the number of initial hydrogen bonds: 15 hydrogen bonds between deca-alanine and water molecules and 12 degenerate hydrogen bonds within deca-alanine. The degeneracy arises because each hydrogen bonded pair within deca-alanine breaks into two hydrogen-bonding sites that are accessible to the water solvent upon the stretching of the peptide. Hence there are only 6 nondegenerate intra peptide hydrogen bonds in the initial helical structures in agreement with the initial averaged values shown in the top and middle panels of Figure 16. As the peptide is stretched, there is a marked difference in the loss of hydrogen bonds between the vacuum and water solvated cases. In vacuum, deca-alanine maintains most of the intrapeptide hydrogen bonds even when it is stretched by as much as 12Å, whereas in solvent, it nearly loses them all by this point. As indicated in the bottom panel, this is is due to the establishment of hydrogen bonds in the solvent which are evidently not available in the vacuum.

In the solvent, the inter-peptide hydrogen bonds appears to exhibit four regimes. The first of these, during the first few Å's of the forced stretch, exhibits a relatively constant number of hydrogen bonds. This is consistent with the energetics in Figure 15 which display a relaxation of the structure from its initial to a minimum energy value during which the structure undergoes intramolecular reorganization without breaking the hydrogen bonds. The second and third regimes correspond to the lost of the first three and last three intrapeptide hydrogen bonds, respectively. The slope of the loss of hydrogen bonds appears to be slower in the second regime than in the third regime and corresponds to the small change in the slopes in the PMF. Finally,

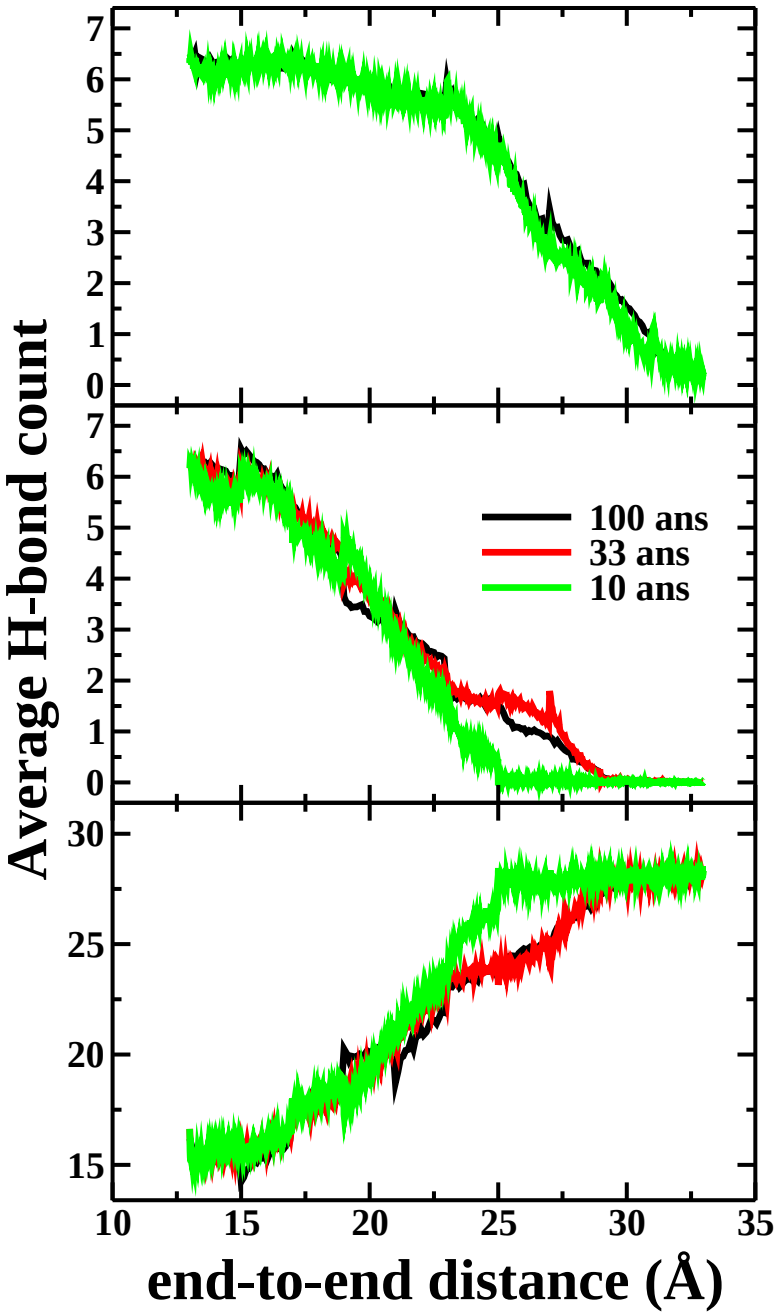

Figure 16: The average number of internal hydrogen bonds in deca-alanine in vacuum and solvent as a function of its end-to-end distance is shown in the top and middle panels, respectively. The average number of hydrogen bonds between deca-alanine and the water molecules is displayed as a function of the reaction coordinate in the bottom panel In all panels, the results are obtained for three different pulling velocities: 100Å/ns (black), 33Å/ns (red) and 10Å/ns (green). The only exception is found in the top panel in which the red curve was not computed, and hence is not shown, as there is clear convergence between the displayed curves.

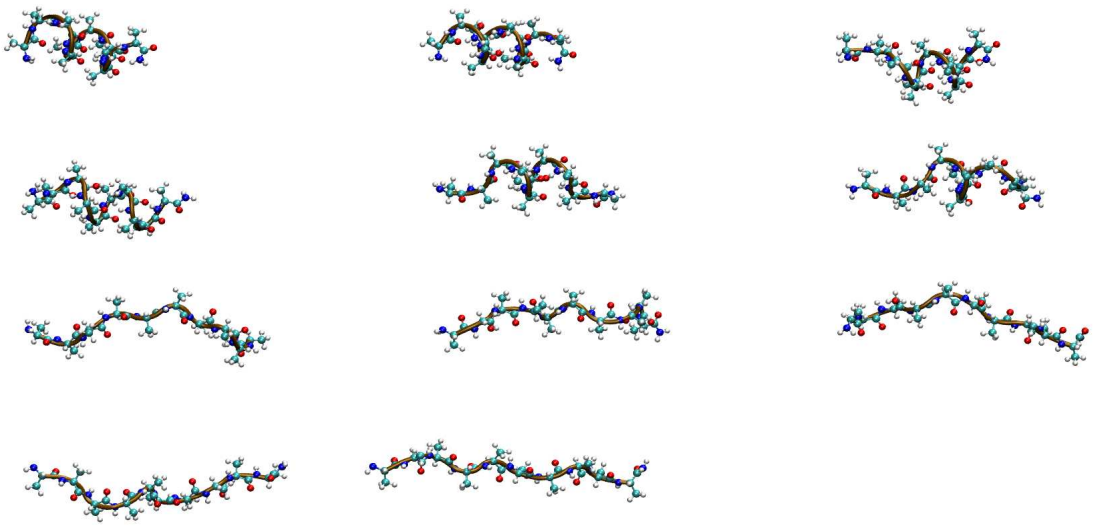

Figure 17: The structure of decaalanine (with the water solvent hidden) is shown at 11 points along a single steered MD unfolding pathway at 300K. The first of these 11 snapshots (top left) is the initial configuration, the remaining 10 snapshots are the final configurations picked at the end of each step according to JA criteria described in this supplementary document. The decaalanine backbone is illustrated using a brown ribbon. Also embedded on the ribbon is the atoms of each alanine residue (blue, nitrogen; cyan, carbon; red, oxygen; white, hydrogen). The reaction coordinate begins from top left and goes towards bottom right by walking along each row.

the fourth regime exhibits little change in hydrogen bonding, and corresponds to the relatively flat energetics in the PMF. These observations are thus consistent with the hypothesis that the energetics of the stretching of NPY are primarily correlated with the making and breaking of hydrogen bonds within the peptide and to the solvent.

4.3.4 The initial configurations in the adaptive SMD stretching of decaalanine in solvent

The structures at the beginning of each step of the adaptive SMD stretching of decaalanine in vacuum are displayed in Figure 17. These structures are selected according to the criteria which selects the structure whose energy best matches the Jarzynski average at the end of the previous step.

4.4 Conclusion

Steered molecular dynamics (SMD) is a non-equilibrium MD method in which a series of external force is applied on a particular atom or group of atoms so that the system proceeds along a desired direction. The nonreversible work required to move the system throughout the simulation is then averaged using the non-equilibrium work relation (Jarzynski's inequality) to calculate the potentials of the mean force (PMF). We have recently introduced an adaptive implementation of this technique where the reaction coordinate is staged into multiple steps [Ozer *et al.*, J. Chem. Theory Comput., 6, 3026-3038 (2010)]. Therein, we have demonstrated that for cases, where the work distribution fluctuates over a large energy range (i. e. unfolding of the neuropeptide Y) and thus large number of trajectories are needed converge the average work, adaptive SMD will restrain the work across each step so that the PMF converges with fewer trajectories.

This chapter demonstrates that adaptive SMD methodology can reproduce the free energy profile of deca-alanine stretching obtained earlier in vacuum [136] using significantly less CPU time. It can also be used to obtain the PMF for the stretching of deca-alanine *in solvent*. In so doing, we also track the number hydrogen bonds within the peptide and between the peptide and solvent. The results provides new insight into the unraveling of a helical peptide and the role of hydrogen bonding therein. Solvent molecules stabilize the stretched deca-alanine by quickly replacing the broken intrapeptide hydrogen bonds. Not surprisingly, the hydrogen bonds that stabilizes the helix of deca-alanine resist the pulling when in vacuum much longer than when the peptide was exposed to solvent.

CHAPTER V

ADAPTIVE STEERED MOLECULAR DYNAMICS OF THE LONG-DISTANCE UNFOLDING OF NEUROPEPTIDE Y

5.1 Overview

The neuropeptide Y (NPY) ligand has been a primary target of many recent pharmacological studies because of its implicated function in the brain [174, 175, 176, 177, 178]. Consisting of 36 amino acids, NPY is the most abundant neuropeptide in the mammalian central nervous system [174] and widely expressed in the peripheral nervous system [175]. Several important physiological activities such as induction and control of food intake, inhibition of anxiety, increase in memory retention, presynaptic inhibition of neurotransmitter release, vasoconstriction and regulation of ethanol consumption have been attributed to NPY [176]. The multifunctionality of NPY is the result of its affinity to bind to at least six receptor subtypes—enumerated as Y1 through Y6—belonging to the rhodopsin-like superfamily of G protein-coupled receptors. It has been shown that receptors Y1, Y4 and Y6 are closely related to each other [177]. A recent study on the evolution of neuropeptide Y receptors (Y3 was not investigated) has lead to a partitioning into three subfamilies of receptors: Y1/Y4/Y6, Y2 and Y5 [178].

NPY is a member of the pancreatic polypeptide (PP) hormone family that includes also pancreatic polypeptide (PP) and peptide YY (PYY) [179]. All three ligands share a common hairpin-like structure in tertiary form called the PP-fold. Therein, the the N-terminal residues (1-8) adopt a polyproline type II helical conformation (tail), residues 9-13 form a loop that allows the tail to fold onto an α-helix (residues

14-31), and the C-terminal residues (32-36) are so flexible that they do not participate in the α-helical conformation (14-31) [180, 181, 182]. NMR studies have shown that NPY adopts a different conformation in dimeric form [183, 184, 185] or when bound to membrane mimetic, dodecylphosphocholine (DPC) micelles [186, 187]. In this particular state, the NPY tail is observed to be destabilized and positioned away from the α-helix. Recently, Bettio *et al* reported, in contrast to earlier reports [180, 181, 182], that at low concentrations monomeric NPY favors a less ordered structure in which the β-turn of NPY is more destabilized [188].

The numerical study described herein aims to provide a dynamical explanation for the mechanism performed by an NPY molecule during its structural transition between the reported open (PDB [39] ID: 1PPT [189]) and closed (PDB ID: 1RON [185]) conformations as shown in Figure 18. Knowledge of the pathway may be of use in the design of ligands to stimulate NPY towards the desired fold in vivo, regulators for the binding of NPY to lipid membranes, and alternative receptors. The present work, in particular, provides some insight on the likely form—PP-fold or free tail—adopted by NPY as it binds to a receptor.

5.2 Model and methods

The work described in this chapter is structured as follows: High temperature MD simulations are used to accelerate the unfolding process and to observe a possible unfolding pathway for said process. The proposed unfolding pathway is investigated using steered molecular dynamics (SMD) simulations. The free energy along this path is generally obtained from the SMD trajectories through the use of Jarzynski's nonequilibrium work relation. Unfortunately, the standard application of this approach did not converge within available computational resources. An auxiliary central result of this work is the development of a stepwise adaptive SMD scheme for the calculation of the free energy along a nonlinear and large-distance pathway,

reviewed in Section 3.4.

5.2.1 Determining unfolding pathway of neuropeptide Y via molecular dynamics at elevated temperatures

The relative dynamics of the α-helix and tail in NPY immersed in a periodic box of water molecules have been simulated using several computational protocols to overcome the long times needed to follow simulations of the folding process. The focus of the simulations is the unfolding of NPY as it is faster than the folding process while still revealing the folding pathway(s). The initial state of the unfolding process —namely, the protein's crystal structure— is also more clearly defined than the structures of the unfolded protein basin, and this offers additional numerical advantages in attempts to map out the pathway [190].

Molecular dynamics simulations have been carried out using the NAMD [171] molecular dynamics integrator with the forces in NPY specified through the CHARMM force field [172]. The water molecules are treated using the TIP3P model, and 13,178 water molecules are included in the cube. A time step of 1 fs has been employed in all simulations. Electrostatic interactions have been calculated through the particle mesh Ewald (PME) method [191]. Solvated structures are initialized by inserting NPY into an appropriately-sized cavity created within an equilibrated neat water box. These are equilibrated at 50K for 5 ps and subsequently heated gradually to the temperature of interest. An NPT equilibration run (at the desired final temperature) is then performed to ensure that the cubic box has a density consistent with 1.0 atm of pressure. Temperature control is realized within the NAMD program by integrating the Langevin equation with the Brunger-Brooks-Karplus (BBK) method which is a natural extension of Verlet integration. This results in an ensemble of structures in which NPY is constrained to its folded state within an equilibrated solvent inside of a cubic box with sides roughly between 70-75Å.

Each member of the ensemble of solvated folded-NPY structures is allowed to

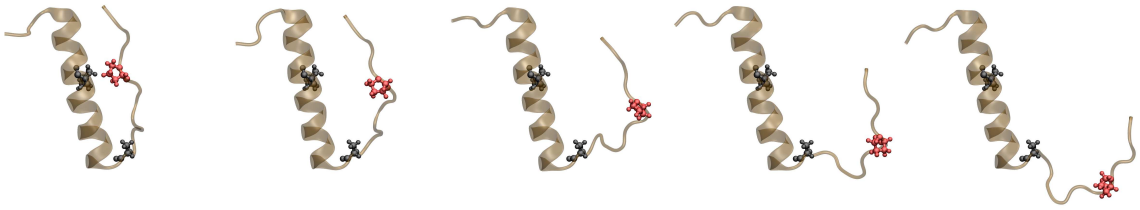

Figure 18: The structure of NPY (with the water solvent hidden) is shown at 5 points along a single steered MD unfolding pathway at 500K. The NPY backbone is illustrated using a brown ribbon. The three residues most clearly marking the unhinging of the tail are shown in atomistic detail: LEU24 (in black on the helix), ALA12 (in black on the turn), and PRO5 (in red on the tail)

freely propagate for 5ps under constant (N, V, E) conditions. It is common practice to run such simulations under constant (N, V, T) conditions using thermostats on all the atoms in the system. However, this has a possible negative side effect of suppressing fluctuations in energy that lead to correlated energy flows between molecules, and more significantly between the unfolding mode and any other mode in the system. The alternative is to run the simulation under constant (N, V, E) conditions at an energy that is thermodynamically consistent with the temperature. This has the disadvantage that the total energy of the box is constant, but with a sufficiently large water box the effective dynamics of the NPY protein will still be that of an open system at constant temperature. The results from a small number of (N, V, T) and (N, V, E) MD simulations are described later, but the conclusion is that all the remaining simulations could be performed using constant (N, V, E) conditions without losing the notion of temperature along the unfolding path.

Although we are primarily interested in the unfolding dynamics of NPY at 310K, the duration of such trajectories is so long that it would entail simulations that are cost-prohibitive. Among several accelerated dynamics approaches now available in the literature, we chose to overcome this obstacle using temperature acceleration [192, 130] as it has been previously reported to accelerate the unfolding process without altering the pathway [127]. Preliminary runs were tested at T=300K, 367K, 433K, and 500K in a cubic box of sides 75Å solvated with equilibrated water (TIP3P) molecules. As will be shown below, NPY unfolded only at 500K within 100 ps, and

hence it became the temperature of choice for the accelerated MD simulations in this work. 500K is well above any natural biological temperature, and is also above the protein melting temperature, T_m. So an experimental system under these conditions may exhibit different dynamics than the biological case [192]. The water system in the computer model, however, remains as a metastable and superheated liquid because neither chemical bond breaking-and-making or evaporation pathways are available to it. The key assumption is that the dynamical pathways also remain in the same universality class, and thus we require additional tests to confirm the predictions of correlation functions using temperature acceleration. As will be discussed below, the model system exhibits the appropriate chemical structures (in the same universality class) as those of the lower temperature.

5.2.2 Steered molecular dynamics of the unfolding of neuropeptide Y

As will be explained in detail during the discussion of the results (i.e. Section 5.3.1), the reaction coordinate of the unfolding of neuropeptide Y is estimated as an unhinging of the polyproline tail away from the α-helix. Steered MD simulation is, therefore, performed by pulling PRO5 at a constant velocity relative to the alpha helix on NPY. The choice of PRO5 is motivated both by experiment and computation. It has been previously reported that 1-4 amino acids of NPY (TYR1 to LYS4) form salt bridges with corresponding receptors [193]. Recent studies have indicated that binding hot spots at protein-protein interfaces exhibit high frequency fluctuation [194]. This suggests that the four residues from TYR1 to LYS4 of the NPY tail fluctuate faster than the other tail residues. Therefore, the choice of PRO5, rather than one of these other residues, allows us to drive the unfolding of the semi-rigid tail (including residues PRO5 to ASP11) while allowing the residues from TYR1 to LYS4 to fluctuate freely. Meanwhile the alpha helix must be represented by at least two fixed points so as to define the requisite hinging motion. These residues are LEU24 on the α-helix and

ALA12 on the hinge connecting the helix to the polyproline tail. The constrained system can therefore be designated through two variables: the LEU24-ALA12-PRO5 angle and the ALA12-PRO5 distance.

5.2.3 Adaptive steered molecular dynamics of the unfolding of neuropeptide Y

As will be seen in Section 5.3.2, the estimated free energy of the extended motion of a finite number of NPY unfolding trajectories is dominated by very few lowest energy trajectories once the applied work gets much greater than a few k_BTs. This results in a weak upper bound to the PMF, which is also observed as the non-vanishing third order term in the cumulant expansion. As discussed in Chapter III thoroughly, adaptive integration of the steered MD methodology overcomes this problem by restraining the generated work distribution within Gaussian range at all points along the reaction coordinate. This is achieved by staging the reaction coordinate into multiple steps. Any given step, i, is sampled over an initial configuration that had been obtained at the end of the previous step, $i-1$. In the case of the unfolding of neuropeptide Y, the unfolding pathway (Section 5.2.2) was divided into 20 steps.

5.2.4 Transition state theory and rates

The experimental results, unfortunately do not provide a potential of mean force that can be used to compare directly to the computational work. Instead, we use the relative stability of the folded and unfolded states (as suggested by the calculated $\Delta G^{\mathrm{u}\leftarrow\mathrm{f}}$) to compare to the experimentally known stable structures. In addition, the rates of the unfolding and folding processes can be determined using transition state theory for the PMF determined along the unfolding path. These will be compared to the findings from both the molecular dynamics trajectories and experiment.

The simple transition state rate is

$$K = \frac{kT}{h} e^{-\frac{\Delta G^{\ddagger}}{kT}} \ , \tag{41}$$

where $\Delta G^{\ddagger}$ is the free energy barrier of the transition. Although much work has been done to go beyond this simple estimate [195, 196, 197], it is reasonably accurate for the order of magnitude of the rate.

5.3 Results and discussion

Analysis of the trajectories was carried out by several methods. Both pepstat, which is our own code, and the NAMD/VMD package were used for trajectory analysis, with the latter focusing on the graphical representations of the trajectories.

5.3.1 Neuropeptide Y unfolds by the unhinging of its polyproline tail away from the α helix

Although the tail section exhibits the most dramatic dynamical changes, structural metrics were collected throughout the protein simulations. Within the polyproline tail (residues 1-12), the time-dependence of the end-to-end distance and radius of gyration,

$$R_g^2 = \frac{1}{N}\sum_{k=1}^{N}(r_k - r_{\text{mean}})^2 \,, \tag{42}$$

are measured. The time-dependence in the tail-to-helix distance is inferred by way of the pairwise distances between residue pairs, 1-31, 4-27, 5-24, 7-20, and 8-16.

The results shown below [cf. Figure 19(b)] suggest that the unfolding pathway involves the unhinging of the tail away from the α-helix instead of sliding. This unhinging occurs about the pivot represented by the ALA12 residue, and is measurable through a so-called tail-turn-helix angle. While the α-helix is relatively stiff through this unfolding, the N-terminal of the polyproline tail—and particularly TYR1 to LYS4—is much floppier. The remaining residues (PRO5 to ASP11) on the tail follow a smoother unhinging and can be used to define the tail-turn-helix angle.

Unfolding of NPY was first investigated through unconstrained MD simulations. MD trajectories were propagated using NAMD with the CHARMM forcefield in an explicit water solvent (TIP3P). A total of 50 independent free MD simulations were

(a) (b)

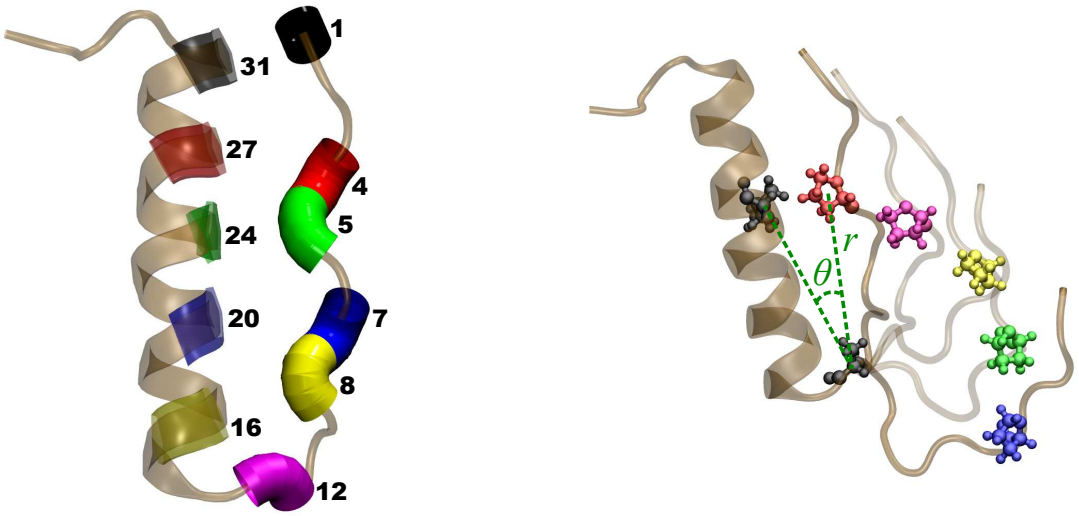

Figure 19: (a) A backbone ribbon diagram of NPY is shown in brown with the helix emphasized by the thick ribbons as usual. The residue, ALA12, at the turn is shown in magenta and acts as the hinge Pairs of residues that are in contact in the folded NPY and whose relative distances and angles are tracked in the following are color coded as in the following scheme; black for TYR1 and ALA31, red for LYS4 and TYR27, green for PRO5 and LEU24, blue for ASN7 and TYR20, and yellow for PRO8 and ASP16. Note that residue positions 14-31 correspond to the helix. (b) The unfolding path is illustrated on the right, wherein the helix and hinge regions are held fixed while 5 images of the tail are overlayed. The PRO5 residue, which is explicitly used for steering relative to the fixed residues LEU24 and ALA12 (shown in black), is shown in five different colors along the unfolding path: red $\rightarrow$ magenta $\rightarrow$ yellow $\rightarrow$ green $\rightarrow$ blue.

performed for this analysis of NPY unfolding at each of several temperatures, 300K, 367K, 433K and 500K. At low temperatures, no unfolding was observed in 1ns simulations, (not shown). At 500K, all of the 50 generated trajectories unfolded within a 1ns observation window.

Detailed analysis of the time-dependence of the helix-tail separation in the 500K unfolding trajectories reveals a hinge-like motion. The distance between the five pairs of residues initially in contact within the folded NPY are shown in the bottom panel of Figure 20. Pairs of residues farther from the turn (ALA12) move to a more distant positions as the protein unfolds. All but the farthest residue pairs sweep a similar angle relative to the turn (ALA12) as shown in the top panel of Figure 20. This suggests that tail hinges away from the helix about the turn during the unfolding process. It does not, however, follow this path linearly. The farthest residue pairs violate the quantitative agreement because the residues at the end of the tail are much more mobile and exhibit large fluctuations in position.

Both experimental [198, 199] and computational [193] studies have suggested that residues 1-4 of the NPY tail form a pharmacophore that plays an active role during NPY binding to receptors. As postulated by Ertekin *et al* [194], interface residues that are in close contact with binding protein residues have a higher packing density and exhibit high frequency fluctuation [194]. The dynamics of the tail shown in Figure 21 is in good agreement with these previous reports. The part of the tail that is proximal to the hinge (including all the residues up to PRO5) have nearly similar geometric properties (in terms of the length and radius of gyration) through the entire unfolding process. The rest of the tail, however, exhibits a significant geometric change through the unfolding process. It appears to be relaxation of the tail end toward a more compact structure in the vicinity of the proximal part of the tail.

The unfolding path thus appears to be primarily following the unhinging of the

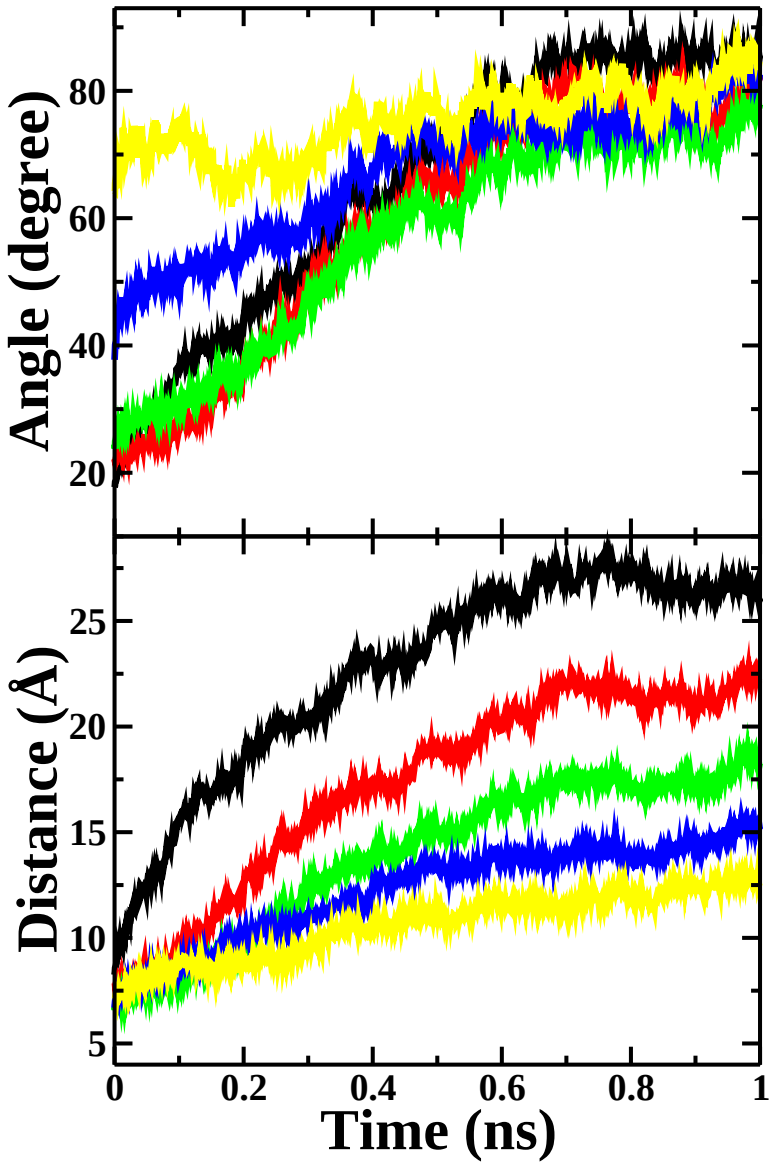

Figure 20: The bottom panel displays the average time-dependent displacement between the tail-helix residue pairs identified in Figure 19 labeled by the same colors as NPY unfolds at 500K. The top panel displays the corresponding time-dependent angles spanned by a given pair of residues with respect to ALA12.

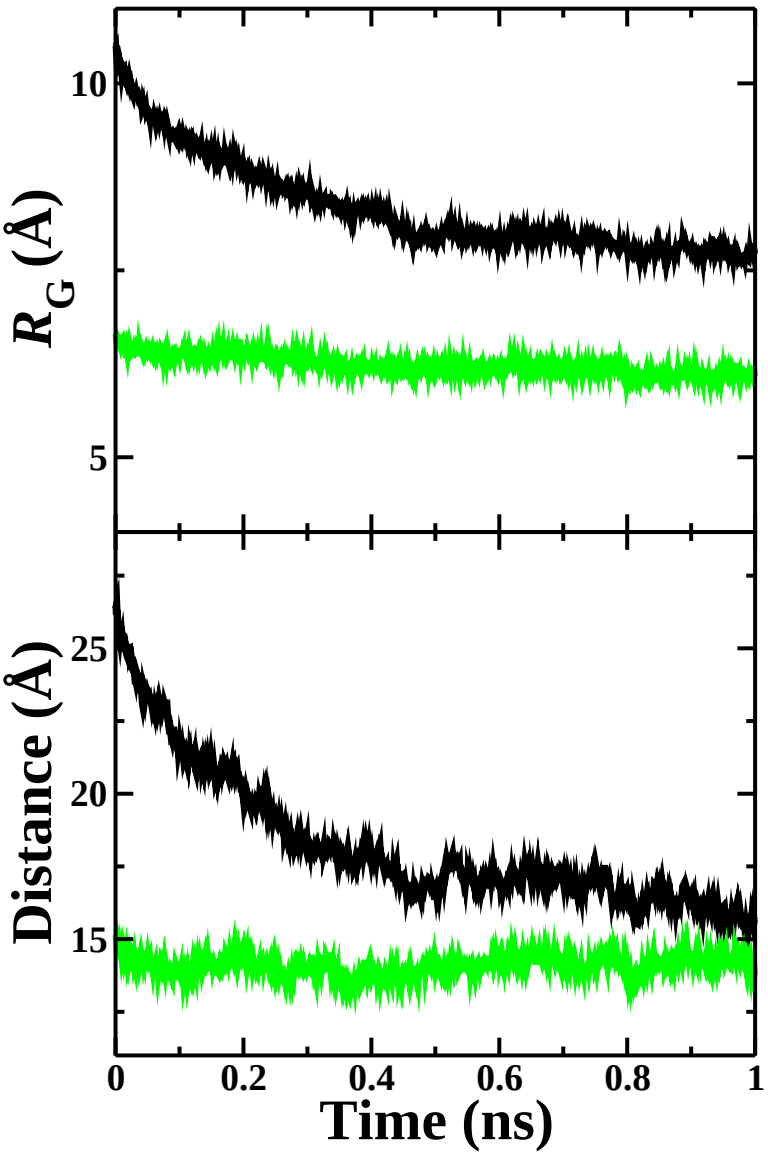

Figure 21: The top and bottom panels display the radius of gyration and end-to-end displacement as a function of simulation time of two different segments of the NPY tail: ALA12 to TYR1 (black) and ALA12 to PRO5 (green). [cf. Figure 19(b) for the identification of the residues.]

proximal part of the tail about the ALA12 hinge. Through this process, the tail appears to be nearly rigid up to PRO5, while it the more distant residues are much more mobile. Hence PRO5 is associated in the remainder of this work with the unfolding (reaction) path illustrated in Figure 20. Following Daggett and coworkers [200], we therefore suppose that this unfolding path is followed not just at the elevated temperature of 500K, but also at experimentally accessible temperatures.

5.3.2 Potentials of mean force obtained using steered molecular dynamics is dominated by the rare low energy configurations

Our objective is to learn about the dynamics of NPY at temperatures relevant to the experimental systems. The accelerated MD simulations provided us rates only at the locally stable temperature of 500K. They also suggested an unfolding path along which we can calculate the PMF at lower temperatures for the purpose of obtaining relative rate information as will be done in the next subsection. The PMF must be calculated at 500K for comparison with the MD simulations. For the lower temperature, we choose 310K as is the so-called body or *in vivo* temperature and is the temperature at which several experimental studies have explored the NPY dynamics [181, 182, 186]. The determination of the PMF at these two temperatures is nontrivial because the models are quite large (consisting of 40,123 atoms) for which a single nanosecond trajectory takes approximately 100 hours on one computer core. Nevertheless, the non-equilibrium SMD approaches described in Chapter III were used to obtain the PMFs. The non-equilibrium simulations were realized using NAMD with the CHARMM forcefield for NPY in an explicit water solvent (TIP3P). All standard configuration parameters were the same as in the unconstrained MD simulations. The PMFs determined by either SMD approach required 110 hours running on forty-eight 2.33GHz Intel 64 CPUS for 144 1ns trajectories at a cost of 5280 CPU hours.

Steered MD trajectories have been obtained at high temperature (500K) as well as at body temperature (310K). The unhinging of the tail was steered by pulling

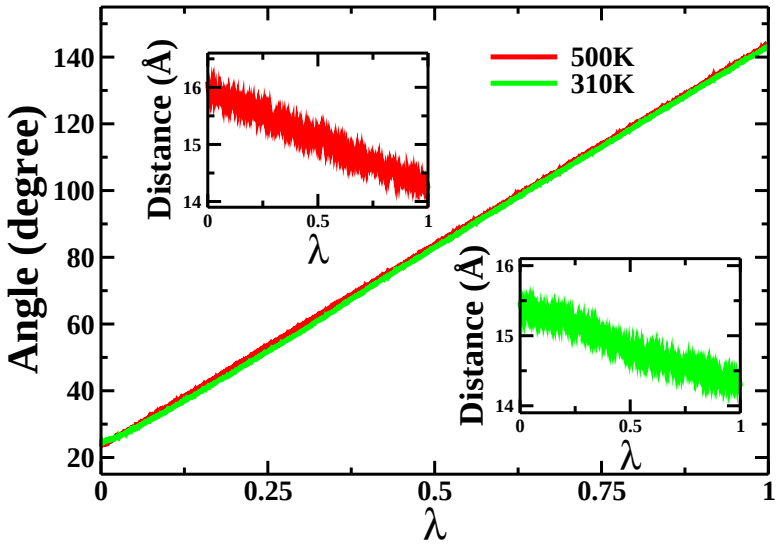

Figure 22: The average displacements in the system are shown along the parameterized path λ for the adaptive SMD at 310K and 500K. The displacements $\xi_x = (r, \theta)$ fixed by the non-equilibrium process correspond to the radial distance r of PRO5 from the hinge (ALA12) and the angle θ spanned by the PRO5-ALA12 and LEU24-ALA12 vectors.

PRO5 (coupled to a dummy atom through a spring constant as per Equation (37)) relative to the virtually fixed residues ALA12 at the turn of the loop and LEU24 on the α-helix. The unfolding path, which the dummy atom follows, is a discretization of the pseudocircular path shown in Figure 19(b) with each of the N finite steps taken to be linear. Specifically, the external force was applied on PRO5 to steer it from an initial configuration of the PRO5-ALA12-LEU24 angle θ_{initial} and radius r_{initial} to the final values, θ_{final} and r_{final}. At 500K, $\theta_{\text{initial}} = 24.36°$ and $r_{\text{initial}} = 16.09$Å. At 310K, $\theta_{\text{initial}} = 24.41°$ and $r_{\text{initial}} = 15.49$Å. At both temperatures, the final configuration is $\theta_{\text{final}} = 144.4°$ and $r_{\text{final}} = 14.3$Å. The initial configurations for the two temperatures differ because each were prepared from equilibration runs at the respective temperatures. All control parameters, such as pulling velocity ($v = 33$Å/ns) and spring constant ($k = 7.2$kcal/mol), were kept identical to each other so as to render comparable results. The degree to which the PRO5 residue followed the unfolding path through the SMD simulations is shown in Figure 22. On average, both

θ and r follow the linear displacement well as expected for a constant velocity pulling SMD simulation. The fluctuations around the average are small and also consistent with this conclusion.

At each temperature, 144 independent SMD trajectories were generated. (The number is 144, not 100, because of technical reasons related to the architecture of the particular computer cluster and the number of simultaneous trajectories—three—that could be run per core without increasing the wall clock time.) This number was sufficient to converge the adaptive SMD trajectories and therefore serves as a good foil for the comparison of the two methods utilizing a similar amount of computational resources. Figure 23 shows the work and the averaged PMF using Jarzynski's relation at both 500K (top) and 310K (bottom). There are only a limited number of trajectories contributing to the PMF of the system at each temperature. This suggests a need for many more trajectories in order to converge the Jarzynski average. Indeed, the original deca-alanine in vacuum SMD PMFs calculated by the Schulten group [153, 136] required over 10,000 trajectories on this much smaller system.

The lack of convergence of this approach (using a limited number of trajectories) is also illustrated by the comparison of the PMF between Jarzynski's average and the second order cumulant expression shown in Figure 24. The two expressions are equal In the limit that the work distribution is Gaussian because of the well-known Marcinkiewicz's theorem [201]. The lack of agreement between the two expressions is due both to the use of too few trajectories and also the fact that the observed trajectories were able to stray far from the relevant configurations. The consequence of the latter is that the statistics of the work contributions are far from Gaussian and hence the second order cumulant expression deviates greatly from Jarzynski's average. As mentioned in Section 3.4, this problem can be overcome by decreasing the pulling velocity. The problem with this approach is that in order to acquire a converged PMF one may need to decrease the pulling velocity as low as the reversible

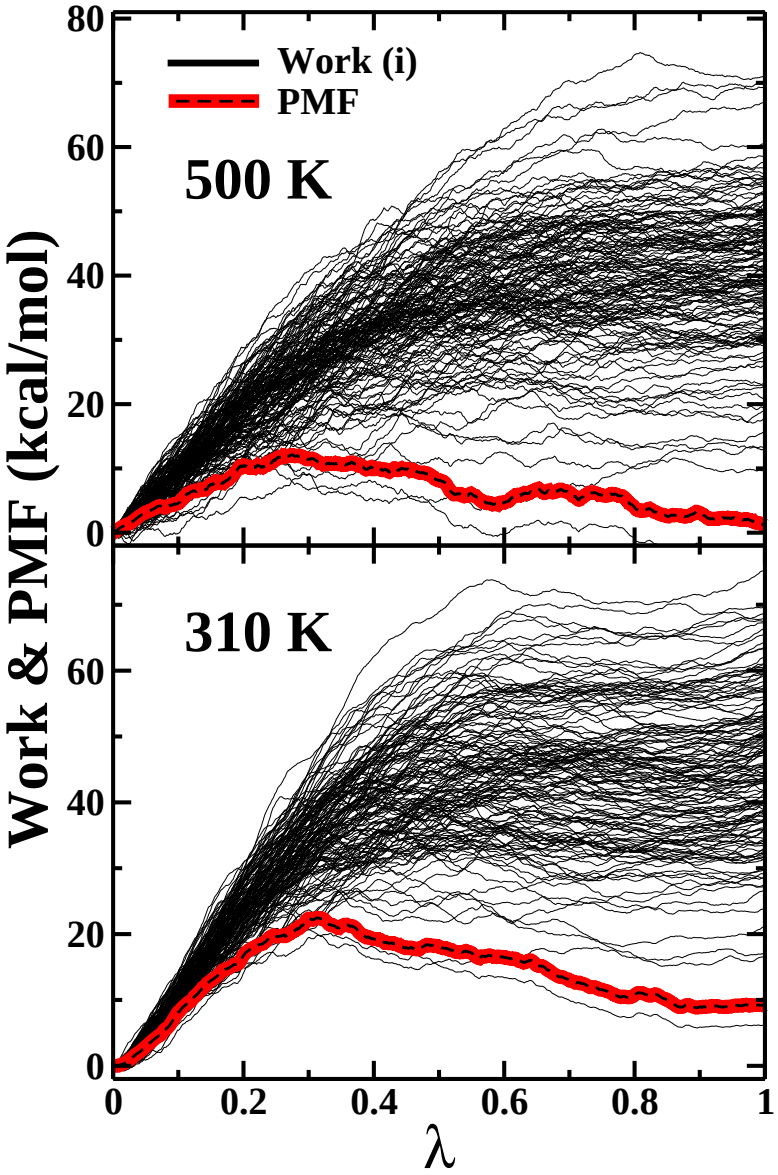

Figure 23: The work for 144 individual trajectories α (in black) and the PMF (in thick-red highlighting of a black-dashed curve) obtained using the Jarzynski equality are displayed as a function of the parameterized unfolding path at 500K (top panel) and 310K (bottom panel).

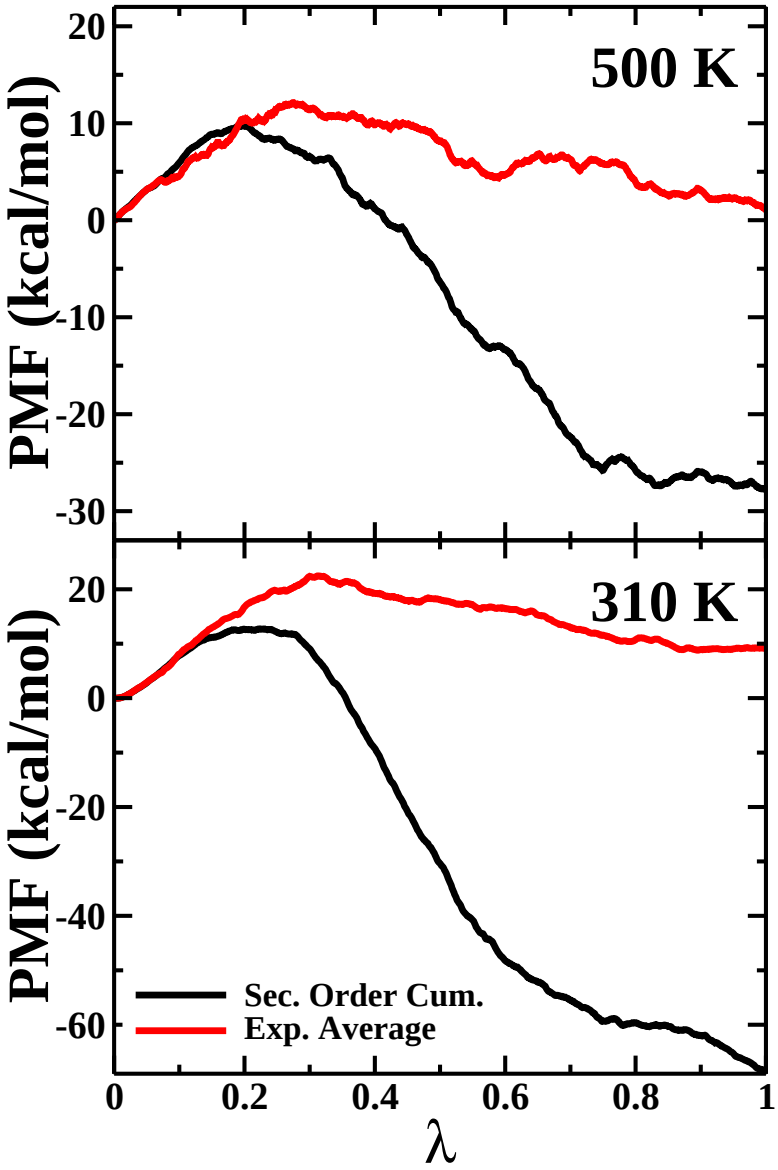

Figure 24: The PMF obtained using Jarzynski's equality (red, cf. Equation (22)) and second order cumulant expression (black, cf. Equation (24)) obtained from a standard SMD calculation with 144 trajectories are displayed as a function of the parameterized unfolding path at 500K (top panel) and 310K (bottom panel).

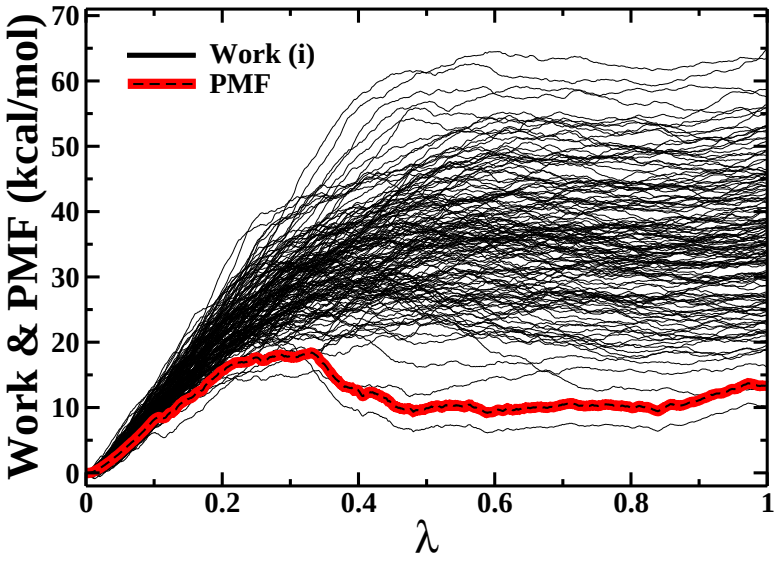

Figure 25: The work for 144 individual trajectories α (in black) and the PMF (in thick-red highlighting of a black-dashed curve) obtained using the Jarzynski equality are displayed as a function of the parameterized unfolding path at 310K. The pulling velocity in this experiment is decreased by 50% compared to the experiment that led to Figure 23.

velocity which is computationally not feasible for a large system such as neuropeptide Y. Figure 25 demonstrates the indifference between the 33Å/ns pulling (Figure 23) and the 17Å/ns pulling (Figure 25). The PMF in the latter is also dominated by the lowest energy trajectory. Although not shown, the second order cumulant of this slower pulling experiment did not converge onto the Jarzynski's average, neither.

5.3.3 Potentials of mean force obtained using adaptive steered molecular dynamics converge with significantly fewer trajectories

The adaptive SMD method described in Chapter III preempts the work distribution of high barrier PMFs from losing their Gaussian nature by partitioning the unfolding path into several steps over which the PMF undergoes smaller changes. For the curved unfolding path illustrated in Figure 19(b), we found convergence when we used 20 steps and a mere 144 trajectories per step. As noted earlier, the total computational cost is almost the same excluding the negligible cost required for trajectory comparison at the end of each step. As before, 144 independent adaptive SMD trajectories

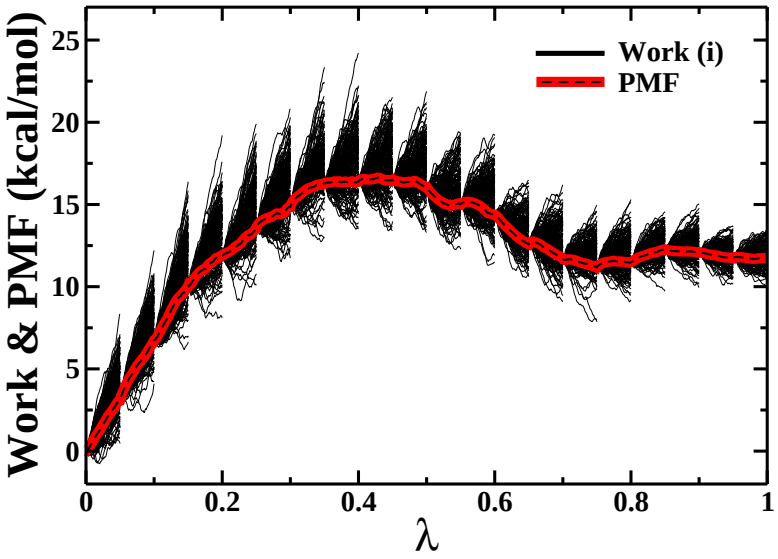

Figure 26: The work for 144 individual trajectories α (in black) and the PMF (in thick-red highlighting of a black-dashed curve) obtained using *adaptive* SMD are displayed as a function of the parameterized unfolding path at 500K.

were generated for each of the two temperatures, 310K and 500K.

The work and the averaged PMF using adaptive SMD (Equation (35)) are shown in Figure 26 (500K) and Figure 27 (310K). Unlike in the results for the standard SMD simulations shown above, the PMFs are not dominated by the lowest energy trajectories. On the contrary, the PMF for each step has contributions from several trajectories. The results obtained for the PMF using the adaptive SMD method (cf. Equation (35)) with Jarzynski's equality (cf. Equation (22)) shown in Figures 26 and 27 are reproduced in Figure 28. Therein, the PMFs obtained with the second order cumulant expression (cf. Equation (24)) are also shown. The agreement is remarkable as the differences are not visible at this level of resolution. Though not shown, the number of sampled trajectories was doubled leading to no significant change in the converged PMFs. Thus the adaptive non-equilibrium process appears to result in a better estimate of the PMF with a limited number of trajectories, i.e., computational resources.

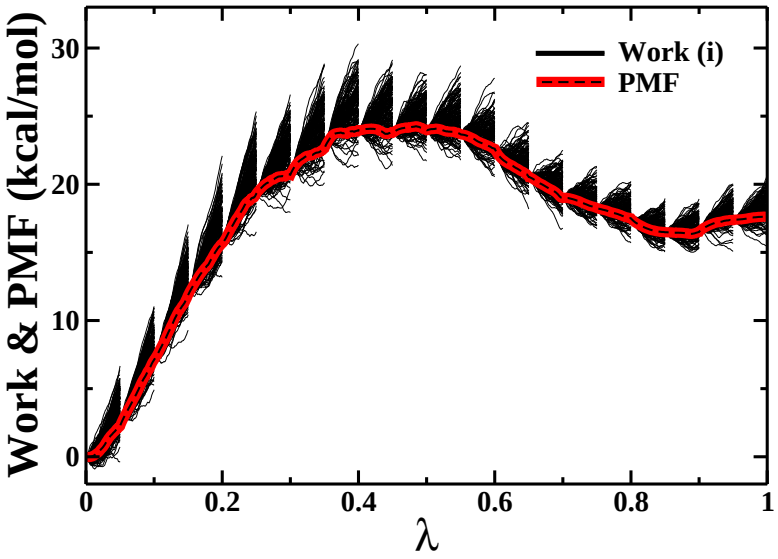

Figure 27: The work for 144 individual trajectories α (in black) and the PMF (in thick-red highlighting of a black-dashed curve) obtained using *adaptive* SMD are displayed as a function of the parameterized unfolding path at 310K.

The PMFs in Figure 28 also provide information about the energetics of the unfolding process of NPY at the two temperatures, 310K and 500K. The barrier height to unfolding is approximately 15 times the elevated temperature whereas that ratio is 40. As NPY had exhibited only partial unfolding at the high temperature, it is therefore not surprising that the low temperature MD simulations did not unfold within the 1 nanosecond observation window. In addition, the folded state has a lower PMF and is therefore predicted to be the more stable form for monomeric NPY at 310K.

5.3.4 The folding and unfolding rates of NPY

The barrier height for the transition from the folded to unfolded conformations of NPY has been found to be 24 kcal/mol and 17 kcal/mol for 310K and 500K, respectively. From these activation energy values, the rates have been calculated as $5.1 \times 10^{-5} s^{-1}$ and $5.5 \times 10^{5} s^{-1}$ again for 310K and 500K, respectively. The inverse of these rates corresponds to a lifetime for the NPY unfolding transition. At the accelerated temperature (500K), this lifetime is 1.8 μs and is consistent with the fact that the NPY trajectories would explore the unfolded space within 1ns as seen in

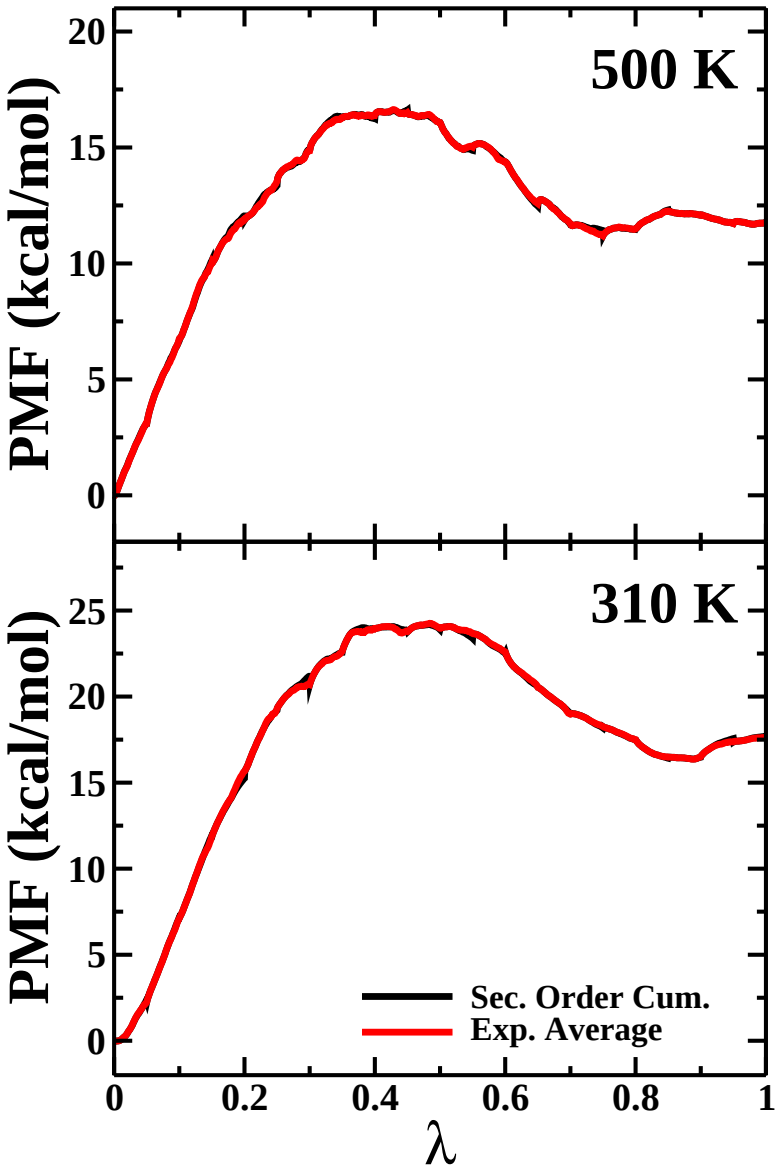

Figure 28: The PMF obtained using Jarzynski's equality (red, cf. Equation (22)) and second order cumulant expression (black, cf. Equation (24)) obtained from an *adaptive* SMD calculation with 144 trajectories are displayed as a function of the parameterized unfolding path at 500K (top panel) and 310K (bottom panel). Note that the black curves are nearly entirely covered by the red curves and hence not very visible.

the MD simulations. At the body temperature (310K), this would suggest a lifetime over 5 hours which is consistent with the fact that none of the low temperature NPY proteins unfolded during the MD simulations.

5.4 Conclusion

The unfolding path of NPY has been suggested by accelerated MD simulations to be the unhinging of the polyproline tail away from the alpha helix about the turn (near ALA12.) The NPY tail maintains its overall shape between PRO5 and ASP11 while unhinging away from NPY helix. As the NPY unfolds along the path, the first four N-terminal residues (TYR1 to LYS4) fluctuate freely when no biasing force is applied on them. This observation is consistent with earlier reports which hypothesized that these fours residues on the polyproline tail of NPY form a pharmacophore at the NPY-receptor interface during NPY bioactivity [193]. This has been justified by the fact that protein-protein interfaces have been seen to be enriched in the presence of high frequency fluctuating residues [194].

The potentials of mean force along the folding path provide a more detailed view of the dynamics. This was possible because of a generalization of SMD (also known as force-biased simulations) using the adaptive scheme introduced in this work. The barrier heights and associated rates of the NPY unfolding transition at an accelerated temperature (500K) and the *in vivo* temperature (310K) agree well with the numerical MD simulations (reported here) and those authors [180, 181, 182] which have proposed the stability of PP-fold based on their experimental findings. At the *in vivo* temperature, we have determined an unfolding rate for NPY on a time scale longer than 5 hours. The typical single-domain protein folding/unfolding time scale is a few μs at the fastest and a couple hundred μs at the slowest [202]. We thus conclude that at 310 K monomeric NPY does not unfold. This conclusion is consistent with our

preliminary unconstrained MD simulations in which NPY did not unfold at temperatures up to 433K. The fact that the unfolded NPY state has a higher free energy than the folded structure also suggests that NPY monomer in solution is folded in the pancreatic-polypeptide (PP) fold. This result is also consistent with the experimental hypothesis that the NPY dimer is biologically inactive in solution because the tail moves away from the PP-fold [182]. This indirectly suggests that the biological activity of the NPY monomer results from the stability of the folded structure in agreement with the energetic stability found in this work. Recently, Bader *et al* [186] reported that micelle-bound form of NPY demonstrates a less ordered conformation than the PP-fold. In this less-ordered conformation, the NPY tail is observed to be fluctuating (Figure 3 in Bader *et al*) while the α-helix remains stable. Our results suggest that this is due to the specific contacts, formed between micelle and side chains of NPY α-helix, replacing the favorable polyproline tail and α-helix contacts observed in the PP-fold.

CHAPTER VI

ADAPTIVE STEERED MOLECULAR DYNAMICS OF THE LONG-DISTANCE UNFOLDING OF PORCINE YY AND VARIOUS MUTANTS OF PORCINE YY

6.1 Overview

Adopting the same pp-fold as neuropeptide Y (NPY), porcine peptide YY (PYY) is another widely studied naturally occurring helical hairpin. PYY assumes the native pp-fold in which a type II polyproline (PPII) helix (at the N terminus) is folded onto an α-helix (at the C-terminus). The stability of the pp-fold is characterized by a well defined hydrophobic cluster emerging from the side chain interactions between the PPII helix and the α-helix [203, 204, 187]. Although the folding mechanism and unfolding thermodynamics of helical hairpins are investigated extensively, most of these studies are based on the kinetic research of the helical hairpins which are stabilized by the disulfide bridges formed between the α-helix and PPII helix [205, 206, 207, 208]. For example, the folding mechanism of Z34C, an example to cystine stabilized helical hairpin, has been proposed recently by Du and Gai [209]. The existence of a very strong disulfide bond between the two helices of a helical hairpin artificially reduces its accessible conformational space. As a result, thermodynamic and kinetic results obtained from studying such hairpins may not reflect their complete folded and unfolded ensemble. It is, therefore, of great significance to elucidate thermodynamics and kinetic information regarding the folding/unfolding mechanism of PYY, which assumes the helical hairpin like pp-fold without any disulfide bridge.

Recently, Waegele and Gai have reported on the kinetic behavior of the unfolding of PYY and several mutant of PYY [210]. Their report has provided important

insight on the kinetic roles of the structural elements of the pp-fold such as (i) the hydrophobic cluster between the two helices and (ii) the stable turn. Their experimental study is based on the comparison of the infrared response displayed by the native PYY and several of the mutants. They have measured the relaxation rate, folding rate and unfolding rate at several temperatures ranging from 293K to 343K using temperature-jump infrared spectroscopy [211] in conjunction with site-directed mutagenesis [212]. Mutations are designed as to (i) identify the effect of hydrophobic interactions between the side chains of the two helices of PYY on the stability of the pp-fold (i.e. A7Y, Y21A, Y27A) and (ii) elucidate the structural significance of the turn region (S13A, P14A). An illustration of the positions of the mutated residues is displayed in Figure 29.

Their findings suggest that, at 303K, mutations that weaken the hydrophobic cluster (Y21A and Y27A) reduce the folding rate while the mutation that strengthen the hydrophobic cluster (A7Y) increase the folding rate. On the other hand, at the same temperature, mutations at the turn region result in a decrease in the folding rate. When the temperature is elevated to 323K, the effect of the hydrophobic deletion mutations reverse in the sense that the mutants —S13A and P14A— become fast folders whereas the folding rate of A7Y structure slightly decreases. The high temperature kinetics may suggest that the unfolded state conformation at high temperature is more compact. This is in fact a plausible explanation since hydrophobic interaction strengthens as temperature increases. The numerical study described in this chapter aims to provide a dynamical explanation to the kinetic interpretation that Waegele and Gai reported at both temperatures.

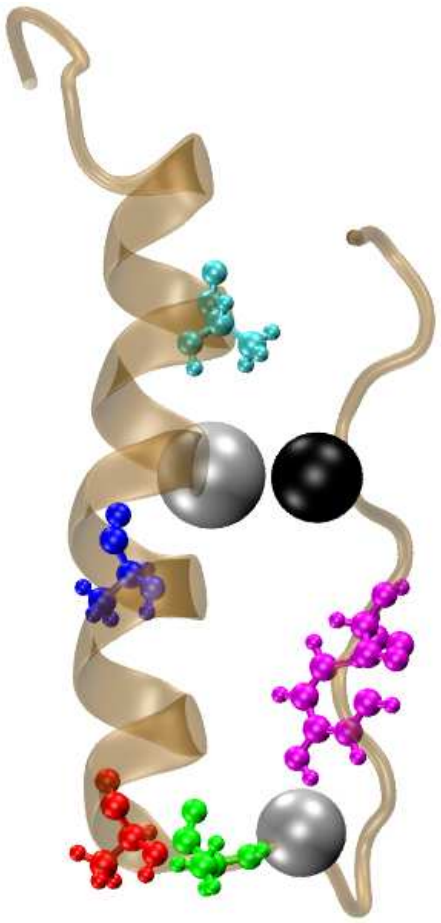

Figure 29: A backbone ribbon diagram of PYY is shown in brown with the helix emphasized by the thick transparent ribbon. Residues that are atomically detailed represent the mutations and they are color coded as in the following scheme; magenta for A7Y, green for S13A, red for P14A, blue for Y21A, and cyan for Y27A. Note that residue positions 14-31 correspond to the helix. The large black dot represents PRO5 which is pulled along a circular unfolding pathway. The large silver dots represent LEU24 on the helix and ALA12 on the turn, respectively. LEU24 is harmonically fixed at the initial position whereas ALA12 is allowed to move freely.

6.2 Model and methods

6.2.1 Adaptive steered molecular dynamics of the unfolding of porcine YY and its mutants

The unfolding pathway of the PYY is assumed to be similar to NPY since they share the same pp-fold in monomeric form. Therefore, the simulation parameters that have been employed to investigate the unfolding of PYY and its five mutants are set to be identical as the parameters employed for the NPY study. The details of the simulation parameters are discussed in Section 5.2.1 of Chapter V. The unfolding pathway of PYY is also assumed to be similar to that of NPY. The only difference is, in the line of the findings of Waegele and Gai, that the turn region is allowed to move freely while in the case of NPY it was constrained with ALA12 fixed. PRO5 has been steered along a circular path to unhinge with respect to the fixed LEU24 and the initial coordinates of ALA12 (not fixed). Adaptive steered molecular dynamics simulations have been implemented along this unfolding coordinate for PYY and its five mutants. PRO5, LEU24 and ALA12 have been marked as dots in Figure 29 for visualization.

6.3 Results and discussion

6.3.1 Potentials of mean force obtained using steered molecular dynamics of porcine peptide YY are not as structured as of neuropeptide Y

Waegele and Gai have reported quantitative rates —at 303K and 323K— for folding and unfolding of PYY and its five mutants. Our objective, is to estimate the activation energy, $\Delta G^{\ddagger}$, from the potentials of mean force calculated via adaptive steered molecular dynamics simulations. The simulations are implemented for each of the six structures at both temperatures. The top panel in Figure 30 displays the PMFs for the unfolding of each structure at 303 K. Although the PMFs do not identify the two state unfolding that had been observed for the case of NPY, except for the A7Y mutant a plateau at fully unhinged state is observed for all of the structures. This,

Table 2: Comparison of experimentally and computationally determined unfolding times at 303 K. Data for the experimental values are reported by Waegele and Gai [210] whereas data for computational values are calculated as the inverse of the TST rates obtained from adaptive steered molecular dynamics simulations.

Structure	(Experimental) k_u^{-1} (μs)	(Adaptive SMD) k_u^{-1} (μs)
PYY0	7.1±1.3	6.1^{7}
A07Y	9.6±1.7	3.5^{8}
S13A	2.8±0.5	1.3^{14}
P14A	9.5±1.7	3.2^{15}
Y21A	4.2±0.8	4.3^{12}
Y27A	2.1±0.4	8.5^{15}

however, is not a long-lived state as the barrier from this unhinged state back towards the folded state is very little. The simulations are also implemented at 323 K. The bottom panel in Figure 30 displays the PMFs for the unfolding of each structure at this temperature. The PMFs again do not show a two state unfolding for the PYY and its mutants.

One can calculate the transition state rates using the difference between the maximum and minimum values of the PMF in order to compare the quantitative folding/unfolding times produced by Waegele and Gai [210]. Table 2 shows the unfolding times obtained from adaptive SMD simulations of the PYY and mutants compared to those reported by their temperature-jump infrared investigation. Since, we have overestimated the potentials of mean force for all structures. Therefore, the activation energies, $\Delta G^{\ddagger}$, inserted in Equation (41) become too high. The resulting rates, thus, are negligibly small to yield almost zero population at the unfolded state.

Same as the results observed at 303 K, the activation energies, $\Delta G^{\ddagger}$, are overestimated. No long-lived unfolded state is observed for all of the six structures analyzed. Table 3 displays the quantitative comparison between the unfolding times observed in our adaptive SMD simulations and those produced by Waegele and Gai. The observation, that the resulting rates are negligibly small at 303 K, is also valid at 323

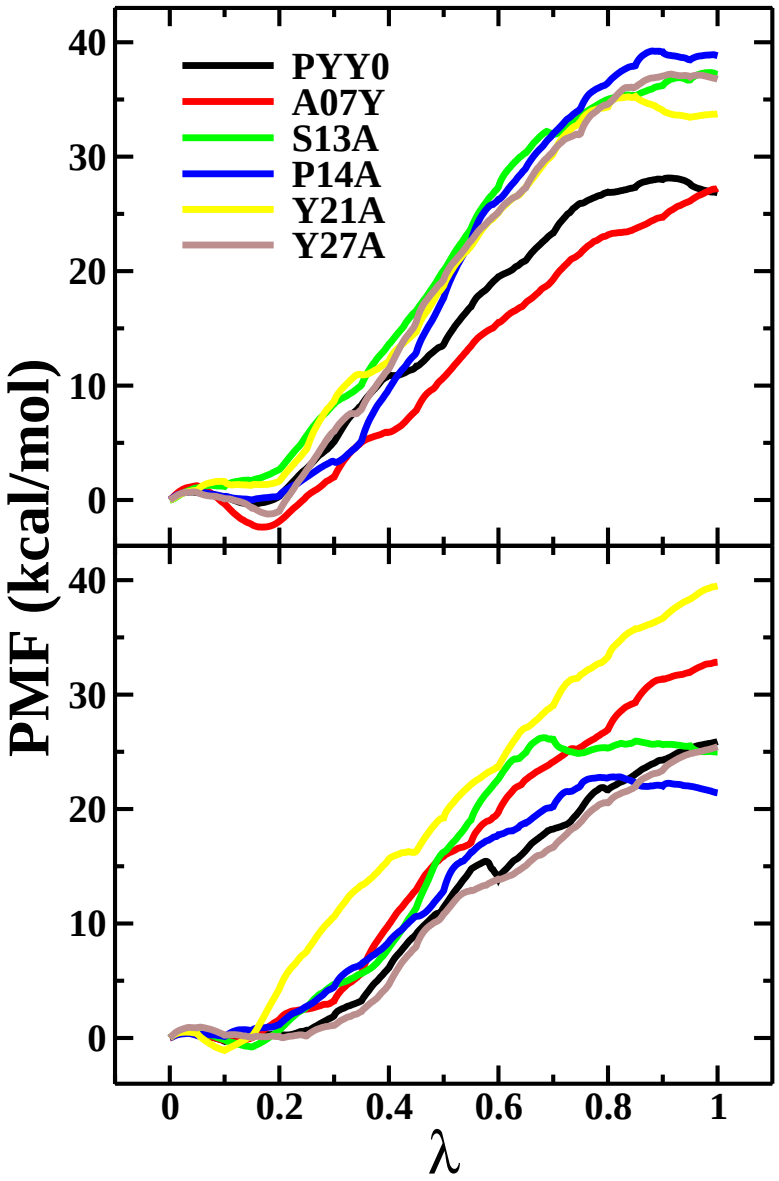

Figure 30: The PMFs —as a function of the parameterized unfolding path— of the unfolding of PYY and its five mutants are generated over an ensemble of 144 trajectories at 303K (top panel) and 323K (bottom panel). Black, the native PYY; red, A7Y; green, S13A; blue, P14A; yellow Y21A; brown, Y27A.

Table 3: Comparison of experimentally and computationally determined unfolding times at 323 K. Data for the experimental values are reported by Waegele and Gai [210] whereas data for computational values are calculated as the inverse of the TST rates obtained from adaptive steered molecular dynamics simulations.

Structure	(Experimental) k_u^{-1} (μs)	(Adaptive SMD) k_u^{-1} (μs)
PYY0	7.1±1.3	6.8^{4}
A07Y	9.6±1.7	4.0^{9}
S13A	2.8±0.5	2.9^{5}
P14A	9.5±1.7	4.1^{2}
Y21A	4.2±0.8	4.3^{14}
Y27A	2.1±0.4	2.4^{4}

K. This is again due to the overestimated potentials of mean force which result in relatively higher values of the transition state activation energy.

One may recall that the unfolding pathway applied on PYY is slightly different than that applied on NPY and argue that different result is simply an artifact of altering the reaction coordinate observed for NPY. Since the unfolding pathway of NPY was not only observed in temperature accelerated MD simulations but also confirmed in adaptive SMD simulations, this is actually a legitimate suspicion. In order to eliminate such suspicion, the adaptive SMD simulations are also implemented on native PYY only along the original unfolding pathway observed for NPY. The results are summarized in Figure 31. The PMFs for the unfolding process where ALA12 was actually fixed display higher values than those for the unfolding process where ALA12 was allowed to move freely. In fact, the estimated activation energies are far too larger than those calculated for the NPY even at 310K (See Fig 27. This indicates that fixing the turn region does not necessarily yield better estimation of the PMFs.

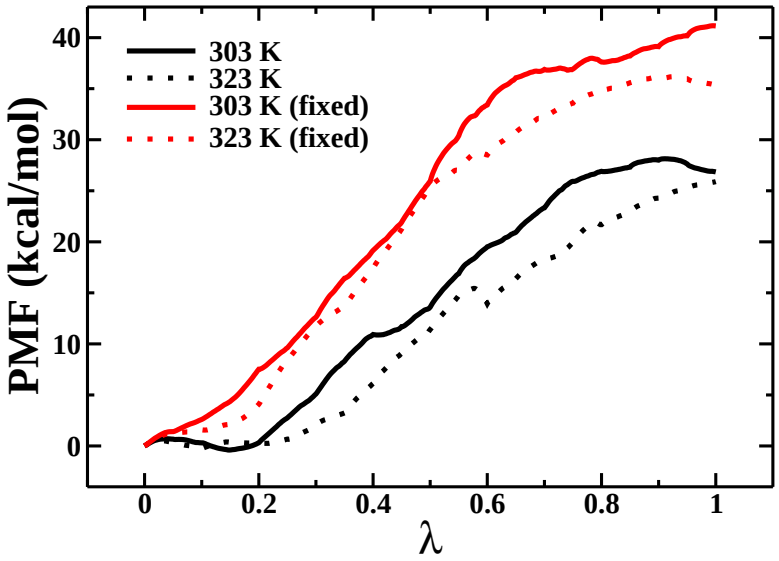

Figure 31: The PMFs —as a function of the parameterized unfolding path— of the unfolding of PYY are generated over an ensemble of 144 trajectories at 303 K and 323K. These PMFs are generated along two slightly different unfolding pathways; (i) PRO5 is pulled along the circular path with ALA12 is not constrained (black straight, at 303 K; black dashed at 323 K) (i) PRO5 is pulled along the circular path with ALA12 is fixed at its coordinate (red straight, at 303 K; red dashed at 323 K)

6.4 Conclusion

The thermodynamics and kinetics of the pp-fold —adopted by the porcine peptide Y (PYY) and several of its mutants— have been investigated using adaptive steered molecular dynamics. Understanding the energetics and kinetics of the pancreatic-polypeptide (pp) fold is of great significance because it is a naturally occurring helical-hairpin in the absence of a dominant disulfide bond. Not being stabilized by the strong disulfide attraction, the stability of the pp-fold can be elucidated by investigating different interactions between the helices and turn region of the hairpin. One way to investigate the roles of specific interactions on the stability is to mutate amino acids and compare the energetics of the mutated peptide to that of the native peptide. In order to understand the effect of hydrophobic interactions between the α-helix and the polyproline helix three single point mutants of the porcine peptide YY (PYY) were created, A7Y, Y21A and Y27A. Similarly, to understand the formation and stability

of the turn region, two single point PYY mutants were created, S13A and P14A. This set of mutations is set to be identical to the experimental design by Waegele and Gai [210] to compare the numerical results obtained from the adaptive SMD simulations of the unfolding of PYY and the above-mentioned mutants to the quantitative rates obtained by Waegele and Gai.

Adaptive SMD simulations have been designed according to the assumption that —since PYY is a member of the same pancreatic polypeptide family adopting the same pp-fold— both should assume the same unfolding pathway. The potentials of mean force obtained from the preliminary simulations do not seem to identify the pp-fold and free tail form of the peptide accurately in the sense that they are not as structured as the PMFs of the unfolding of NPY. It has been observed that slight change in the unfolding pathway defined may lead to dramatic changes in the structure and energetic if the PMFs (Figure 31). This observation shows that we have not yet precisely describe the unfolding pathway of the porcine YY and thus always estimating inaccurate potentials of mean force. Therefore, altering the reaction coordinate applied through adaptive SMD simulations towards the correct path may definitely characterize the expected two state energy landscape of PYY's conformational space.

CHAPTER VII

CONCLUDING REMARKS AND OUTLOOK

7.1 Comparative modeling point of view

D_2Check utilizes the ψ_i-ϕ_{i+1} dihedral angle distributions in addition to traditional ϕ_i-ψ_i (Ramachandran) distributions of a reference set of structures —typically taken to be a well-defined subset of the PDB— to obtain indirect information about the relative propensity for secondary structure of a given protein or residue in comparison with the reference set. The checking function is defined through the relative probability that a given structure in the reference set will have a particular value of the structure entropy of its dihedral angles. The latter, in turn, is defined using a Shannon entropy based on the joint probability distributions for the angle pairs, ψ_i-ϕ_{i+1} and ϕ_i-ψ_i, along the entire structure. D_2Check has previously been shown to provide explicit information about pair residue interactions as well as an indirect picture of long range correlations in a given protein structure by Hernandez and coworkers [84]. It provides different structural information about a protein than that obtainable using other checking functions such as PROCHECK [72, 73] and WhatCHeck [17] and is therefore appropriate as a complementary tool.

Within the framework of this thesis, D_2 analysis has been extended from the protein level to the residue level. Residue level D_2 compatibility of a given protein structure is based on identifying typical and atypical values of amino acids in a given structure. Extensive data mining on the 7,699 distinct structures in the protein data bank showed that the D_2 distribution forms a near perfect gaussian distribution about 0 with a range from -3 t o+3. The distribution leaves only a small fraction (i.e. 0.22%) of the experimentally determined structures outside of this range—atypical

structures. A given protein is among the 17 observed atypical structures only if most of its amino acids have unusual dihedral angles (i.e. absolute value of the residual D_2 score greater than 3). This observation is used to develop a compact graphical representation, color strip, by projecting the calculated residue level D_2 scores on a color coded scheme. In this scheme, red, green, blue corresponds to residual D_2 scores -3, 0 and $+3$, respectively. As hypothesized, the color strip is shown to differentiate the atypical structures (red or blue rich strips) from the typical structures (green rich strips). Color strips, thus, can be used for visual assessment of predicted or experimentally determined structures at a glance.

The work is further extended towards analyzing structural similarities/differences amongst protein families and structural effects of mutations using the difference of two color strips. Graphical representation of color strip difference has been found to be effective in identifying the conformational deviations observed in a family of mutants (several LYS116 mutants of staphylococcal nuclease) with respect to the wild type staphylococcal nuclease. Residue level D_2 analysis has successfully identified the mutation region as well as other active sites of staphylococcal nuclease [148, 149]. D_2 does not claim to spot active sites in a given protein or to grade structural fidelity of a given structure. It does however, provide additional long range information obtained from ψ_i-ϕ_{i+1} angle pairs. In doing so, it might claim a complementary role to the assessment tools that only uses ϕ_i-ψ_i angle pairs.

The D_2Check server enables users to easily and quickly obtain the D_2Check values for a structure that has already been deposited into the PDB or for a new structure, uploaded by the user, at the residue and protein scales. As discussed in previous work [84], these values indicate the degree to which the propensity for secondary structure along a chain is or is not in agreement with the training set from the PDB. This in turn can sometimes identify possible errors in the structure. More importantly, it can also identify regions of the protein that are highly sensitive to tertiary (or

higher) structural elements. Thus it is hoped that this new tool will be useful to the community as it stands.

7.2 Dynamical perspectives

The stability and function of proteins hinge on the relative free energies of contacts within a chain and contacts between its residues and the solvent. Several computational schemes are being developed in order to better characterize both of these types of contact free energies for systems described at the molecular scale—viz. so-called physics-based models. This includes, but is not limited to, replica exchange MD [128], adaptive biasing force MD [131], and free energy perturbation MD [157, 158]. Among the biased integration methods, steered molecular dynamics (SMD) in combination with the Jarzynski's nonequilibrium work relation [134, 135] has been shown to accurately predict free energy profile of bioprocesses along a predefined steering path such as an unfolding coordinate. The computational costs of these approaches, however, is sufficiently large that only a few proteins have been analyzed in vacuum let alone in solvent.

Although SMD, compared to unconstrained MD, effectively reduces the processing cost in modeling large conformational changes of biomolecular systems, the amount of force applied (ranging typically from 500 pN to several thousands pN) is far larger than that applied in AFM experiments (up to a couple hundred pN) from which SMD aims to reproduce the results. As the applied force (thus work) reads larger values, the distribution of work gets distorted from Gaussian behavior which ultimately causes the calculated PMF to be dominated by the lowest energy trajectories. These shortcomings can be surmounted but at significant computation cost by increasing the size of the ensemble—requiring on the order of millions of realizations, by lowering the pulling velocity or by equilibrating the system often and long. Instead, an adaptive

algorithm [143] has been developed extending the Schulten-Jarzynski steered molecular dynamics method for the calculation of PMFs when the subsystem is dragged across long nonlinear paths. In such cases, the PMF can span many $k_{\mathrm{B}}T$'s leading to the sampling of non-equilibrium trajectories with work functions that fluctuate over a very large energy range. Consequently, only a small fraction of the trajectories generated from the SMD contribute nontrivially to the Jarzynski average. In order to numerically converge this average one then needs to generate a large number of trajectories which can be cost prohibitive. The adaptive algorithm allows one to break up the SMD calculation in a series of steps. The free energy difference across each such step is much smaller, and thereby allows convergence of the Jarzynski average with significantly fewer trajectories. In this sense, the adaptive algorithm is not formally better than the standard approach, but it is significantly more numerically efficient.

The conventional SMD methodology using the Jarzynski's equality has previously been implemented experimentally in molecular force pulling experiments by several groups such as RNA unfolding via atomic force microscopy [213], mechanical unfolding and refolding of proteins and nucleic acids via force-measuring optical tweezer experiments [141, 214, 215, 216], and macroscopic mechanical oscillator in contact with a heat reservoir [142] with the underlying theory having been recently clarified by Zimanyi and Silbey [152]. Adaptive SMD could also be extended to such molecular force pulling experiments. Instead of using single constant velocity force pulling, the adaptive procedure would suggest the use of staged (or stepped) force pulling events. The pauses between the stages need only be held long enough so that the environment to the constrained system can relax (while applying zero work). The numerical success of the adaptive technique developed here may find numerous computational applications (i.e. any SMD design can be converted to the adaptive SMD algorithm).

7.2.1 Helix-coil transformation of decaalanine

Adaptive SMD has been implemented to investigate the helix-coil transition of deca-alanine in vacuum. Park and Schulten's data [136] of the potentials of mean force of the deca-alanine stretching has been reproduced utilizing much lower computational cost. The use of as few as 800 trajectories per step has been found to be enough for the calculated PMF to reproduce the PMF obtained using 10,000 standard SMD trajectories. The helix coil transition of deca-alaine in solvent (TIP3P) has also been studied using the adaptive SMD methodology. As expected, water molecules are observed to stabilize the unfolding process by immediately replacing the broken intra-peptide hydrogen bond. This behavior is confirmed using a hydrogen bond count analysis. In vacuum, intra-peptide hydrogen bonds tend to resist forced stretching In solvent, however, intra-peptide hydrogen bonds are broken rather easily as the donor and acceptor favorably interact with water molecules in their vicinity.

Within the context of decaalanine stretching, different selection criteria at each step have also been investigated. These include choosing (i) the trajectory that is closest to the Jarzynski's average (JA), (ii) the minimum energy trajectory (MW), (iii) the configuration that is nearest to the target steering path (RC). It has been shown that RC causes strong energetic fluctuations and does not yield converged PMF. MW underestimates the PMF even lower than the exact PMF obtained from reversible pulling simulations. JA, on the other hand, results in the most robust and converged PMF.

Another important observation from the study of the helix coil transformation of decaalanine is the different energetics of the transition in vacuum with respect to the transition in solvent. The comparison between the PMFs of the stretching of decaalanine obtained in the two media provides substantial insight on the role of the solvent in the helix-coil transformation of decaalanine. The free energy cost to unraveling the chain drops from circa 23 kcal/mole do circa 14 kcal/mole. The first half

of the forced stretch in vacuum leads primarily to deformation of the intramolecular contacts. Alternatively, this stretch in solvent leads to the breaking of intrapeptide hydrogen bonds. The latter are replaced by hydrogen bonds to the solvent which accounts for the energy stabilization described above. This suggests that the unraveling of decaalanine—whether stretched by an AFM or optical tweezers—should be sensitive to the hydrogen-bonding character of a solvent to a measurable degree.

7.2.2 Unfolding of pp-fold neuropeptides

The adaptive SMD algorithm has been applied to the unfolding of neuropeptide Y (NPY). By definition, SMD needs a "reaction coordinate" along which the system will be steered. The unfolding pathway of NPY has been characterized through high temperature accelerated MD simulations (at lower temperature NPY did not unfold within the 1ns simulation window). Unconstrained MD simulations at 500K have resulted in rapid unfolding of NPY via a unhinging like mechanism. Although controversial [192], the temperature jump is claimed to speed up the unfolding of proteins without altering the pathway. Adaptive steered molecular dynamics has fit in at this point to confirm the proposed unfolding pathway at lower temperature (i.e. body temperature). The two state unfolding mechanism of this pp-fold peptide has been observed at the end of adaptive steered molecular dynamics simulations. Finally, the time scale of structural stability of NPY is obtained by way of a determination of the transition state theory rates on the computed surfaces. Through transition state rate calculations, monomeric NPY is shown to be much more stable when the tail is interacting with the helix via side chains (i.e. pp-fold) So, the pp-fold is favored in monomeric form of NPY, which was previously proposed by Nordmann *et al* [180, 181, 182]. Elucidating an accurate folding/unfolding mechanism for neuropeptide Y is essential as the knowledge of the mechanism may lead to voluntary inhibition or stimulation. The contribution towards understanding the structural events during

NPY folding in monomeric form is thus essential as well.

Another test of the adaptive SMD methodology has been implemented to study the unfolding of porcine YY (PYY) and several of its mutants. The significance of studying this system is twofold: (i) PYY is a naturally occurring helical hairpin with no disulfide bridge stabilizing the pp-fold. Accurate and complete sampling of the energy landscape of the unfolding of PYY is therefore very important since any artificial artifact due to the disulfide bridge is avoided. (ii) Recently Waegele and Gai reported an extensive characterization of the folding/unfolding/relaxation kinetics of PYY and five of its mutants at various temperatures. The single point mutations they have described target both the hydrophobic cluster —A7Y, Y21A, Y27A— and the turn region —S13A and P14A— of the porcine YY. The kinetic data they have presented can be used to interpret the role of the hydrophobic interactions between the α-helix and the polyproline helix as well as the formation and stability of the turn region on PYY's folding behavior. They have quantitatively calculated the rates of folding, unfolding and relaxation of porcine YY and the above-mentioned mutants. This means that experimental data is available to compare the computational data obtained via adaptive steered molecular dynamics methodology. Preliminary results do not accurately predict a two state kinetic model for the unfolding of PYY that was observed for the unfolding of neuropeptide Y. However, altering the unfolding pathway even slightly (such as fixing the ALA12 on the turn region) appears to alter the potential of mean force.

Understanding the folding dynamics of the neuropeptide Y is extremely important due to its biological significance. NPY is reported to induce and control food intake, inhibit anxiety, enhance memory retention, regulate neurotransmitter release and ethanol consumption [174, 175, 176, 177, 178]. In a study on the obese (fa/fa) Zucker rats, NPY was shown to have increased secretion in their hypothalamic paraventricular nuclei [217]. Recently, Toretsky and coworkers indicated the key role of

the neuropeptide Y in Ewings sarcoma growth [218], and growth and vascularization of neural crest-derived tumors [219]. Therefore, an increased understanding of how NPY changes conformation between the biologically active and inactive states may contribute to design of ligands to stimulate NPY towards the desired fold and thus regulate its function so as to treat obesity or the growth of cancer cells.

REFERENCES

[1] list of authors, A. and their affiliations appears in the Supplementary Information, Finishing the euchromatic sequence of the human genome, *Nature*, vol. 431, pp. 931–945, 2004.

[2] Levinthal, C., *How to Fold Graciously*, pp. 22–24. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois.: University of Illinois Press, 1969.

[3] Levinthal, C., Are there pathways for protein folding?, *J. Chim. Phys.*, vol. 65, pp. 44–45, 1968.

[4] Anfinsen, C. B., Principles that Govern the Folding of Protein Chains, *Science*, vol. 181, pp. 223–230, 1973.

[5] Zhong, L. and WC Johnson, J., Environment affects amino acid preference for secondary structure, *Proc. Natl. Acad. Sci. USA*, vol. 89, pp. 4462–4465, 1992.

[6] Cohen, B., Presnell, S., and Cohen, F., Origins of structural diversity within sequentially identical hexapeptides, *Protein Sci.*, vol. 2, pp. 2134–2145, 1993.

[7] Han, K. F. and Baker, D., Global properties of the mapping between local amino acid sequence and local structure in proteins, *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 5814–5818, 1996.

[8] Macdonald, J. R. and W. Curtis Johnson, J., Environmental features are important in determining protein secondary structure, *Protein Sci.*, vol. 10, pp. 1172–1177, 2001.

[9] Costantini, S., Colonna, G., and Facchiano, A., Amino acid propensities for secondary structures are influenced by the protein structural class, *Biochem. Biophys. Res Commun.*, vol. 342, pp. 441–451, 2006.

[10] Costantini, S., Colonna, G., and Facchiano, A., PreSSAPro: a software for the prediction of secondary structure by amino acid properties, *Comput. Biol. Chem.*, vol. 31, pp. 389–392, 2007.

[11] Momen-Roknabadi, A., Sadeghi, M., Pezeshk, H., and Marashi, S., Impact of residue accessible surface area on the prediction of protein secondary structures, *BMC Bioinformatics*, vol. 9, p. 357, 2008.

[12] Zhu, Z. and Blundell, T., The use of amino acid patterns of classified helices and strands in secondary structure prediction, *J. Mol. Biol.*, vol. 260, pp. 261–276, 1996.

[13] Adamczak, R., Porollo, A., and Meller, J., Combining prediction of secondary structure and solvent accessibility in proteins, *Proteins*, vol. 59, pp. 467–475, 2005.

[14] A, L. and Marashi, S., Addition of contact number information can improve protein secondary structure prediction by neural networks, *Excli J.*, vol. 8, pp. 66–73, 2009.

[15] Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., and Voelz, V. A., The protein folding problem: when will it be solved?, *Curr. Opin. Struct. Biol.*, vol. 17, pp. 1–5, 2007.

[16] Branden, C. I. and Jones, T. A., Between objectivity and subjectivity, *Nature*, vol. 343, pp. 687–689, 1990.

[17] Hooft, R. W. W., Vriend, G., Sander, C., and Abola, E. E., Errors in protein structures, *Nature*, vol. 381, p. 272, 1996.

[18] Abola, E. E., Bairoch, A., Barker, W. C., Beck, S., Benson, D. A., Berman, H., Cameron, G., Cantor, C., Doubet, S., Hubbard, T. J. P., Jones, T. A., Kleywegt, G. J., Kolaskar, A. S., Van Kuik, A., Lesk, A. M., Mewes, H. W., Neuhaus, D., Pfeiffer, G., Ten-Eyck, L. F., Simpson, R. J., Stoesser, G., Sussman, J. L., Tateno, Y., Tsugita, A., Ulrich, E. L., and Vliegenthart, J. F. G., Quality control in databanks for molecular biology, *BioEssays*, vol. 22, pp. 1024–1034, 2000.

[19] Chothia, C. and Lesk, A. M., The relation between the divergence of sequence and structure in proteins, *EMBO J.*, vol. 5, pp. 823–826, 1986.

[20] Grant, A., Lee, D., and Orengo, C., Progress towards mapping the universe of protein folds, *Genome Biol.*, vol. 5, p. 107, 2004.

[21] Crippen, G. and Maiorov, V., How many protein-folding motifs are there, *J. Mol. Biol.*, vol. 252, pp. 144–151, 1995.

[22] Leonov, H., Mitchell, J., and Arkin, I., Monte Carlo estimation of the number of possible protein folds: Effects of sampling bias and folds distributions, *Proteins*, vol. 51, pp. 352–359, 2003.

[23] Chothia, C., Proteins - 1000 families for the molecular biologist, *Nature*, vol. 357, pp. 543–544, 1992.

[24] Orengo, C., Jones, D., and Thornton, J., Protein superfamilies and domain superfolds, *Nature*, vol. 372, pp. 631–634, 1994.

[25] Wolf, Y., Grishin, N., and Koonin, E., Estimating the number of protein folds and families from complete genome data, *J. Mol. Biol.*, vol. 299, pp. 897–905, 2000.

[26] Govindarajan, S., Recabarren, R., and Goldstein, R., Estimating the total number of protein folds, *Proteins*, vol. 35, pp. 408–414, 1999.

[27] Govindarajan, S. and Goldstein, R., Why are some protein structures so common?, *Proc. Natl. Acad. Sci. USA*, vol. 93, pp. 3341–3345, 1996.

[28] England, J., Shakhnovich, B., and Shakhnovich, E., Natural selection of more designable folds: A mechanism for thermophilic adaptation, *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 8727–8731, 2003.

[29] Li, H., Tang, C., and Wingreen, N., Designability of protein structures: A lattice-model study using the miyazawa-jernigan matrix, *Proteins*, vol. 49, pp. 403–412, 2002.

[30] Shahrezaei, V. and Ejtehadi, M., Geometry selects highly designable structures, *J. Chem. Phys.*, vol. 113, pp. 6437–6442, 2000.

[31] Zhang, Y. and Skolnick, J., Scoring function for automated assessment of protein structure template qulaity, *Proteins*, vol. 57, pp. 702–710, 2004.

[32] MacArthur, M. W. and Thornton, J. M., Conformation analysis of protein structures derived from NMR data, *Proteins*, vol. 17, pp. 232–251, 1993.

[33] MacArthur, M. W., Laskowski, R. A., and Thornton, J. M., Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy, *Curr. Opin. Struct. Biol.*, vol. 4, pp. 731–737, 1994.

[34] Laskowski, R. A., MacArthur, M. W., and Thornton, J. M., Validation of protein models derived from experiment, *Curr. Opin. Struct. Biol.*, vol. 8, pp. 631–639, 1998.

[35] Kleywegt, G. J., Validation of protein models from C^{α} coordinates alone, *J. Mol. Biol.*, vol. 273, pp. 371–376, 1997.

[36] Kleywegt, G. J., Validation of protein crystal structures, *Acta Cryst.*, vol. D56, pp. 249–265, 2000.

[37] Baker, D. and Sali, A., Protein structure prediction and structural genomics, *Science*, vol. 294, pp. 93–96, 2001.

[38] Bradley, P., Misura, K. M. S., and Baker, D., Toward High-Resolution de Novo Structure Prediction for Small Proteins, *Science*, vol. 309, pp. 1868–1871, 2005.

[39] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E., The Protein Data Bank, *Nucl. Acids Res.*, vol. 28, pp. 235–242, 2000.

[40] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D., Basic local alignment search tool, *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.

[41] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucl. Acids Res.*, vol. 25, pp. 3389–3402, 1997.

[42] Li, W., Jaroszewski, L., and Godzik, A., Clustering of highly homologous sequences to reduce the size of large protein database, *Bioinformatics*, vol. 17, pp. 282–283, 2001.

[43] Li, W., Jaroszewski, L., and Godzik, A., Tolerating some redundancy significantly speeds up clustering of large protein databases, *Bioinformatics*, vol. 18, pp. 77–82, 2002.

[44] Li, W. and Godzik, A., Clustering of highly homologous sequences to reduce the size of large protein database, *Bioinformatics*, vol. 22, pp. 1658–1659, 2006.

[45] Soding, J., Protein homology detection by HMM-HMM comparison, *Bioinformatics*, vol. 21, pp. 951–960, 2005.

[46] Colovos, C. and Yeates, T., Verification of protein structures: patterns of nonbonded atomic interactions, *Protein Sci.*, vol. 2, pp. 1511–1519, 1993.

[47] Pontius, J., Richelle, J., and Wodak, S. J., Deviations from standard atomic volumes as a quality measure of protein crystal structures, *J. Mol. Biol.*, vol. 264, pp. 121–136, 1996.

[48] Kanehisa, M. and Bork, P., Bioinformatics in the post-sequence era, *Nature Genetics*, vol. 33, pp. 305–310, 2003.

[49] Ginalski, K., Grishin, N., AGodzik, and Rychlewski, L., Practical lessons from protein structure prediction, *Nucl. Acids Res.*, vol. 33, pp. 1874–1891, 2005.

[50] Ginalski, K., Comparative modeling for protein structure prediction, *Curr. Opin. Struct. Biol.*, vol. 16, pp. 172–177, 2006.

[51] Kryshtafovych, A. and Fidelis, K., Protein structure prediction and model quality assessment, *Drug Discov. Today*, vol. 14, pp. 386–393, 2009.

[52] Marti-Renom, M., Stuart, A., Fiser, A., Sanchez, R., F, F. M., and Sali, A., Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.*, vol. 29, pp. 291–325, 2000.

[53] Melo, F. and Sali, A., Fold assessment for comparative protein structure modeling, *Protein Sci.*, vol. 16, pp. 2412–2426, 2007.

[54] Fasnacht, M., Zhu, J., and Honig, B., Local quality assessment in homology models using statistical potentials and support vector machines, *Protein Sci.*, vol. 16, pp. 1557–1568, 2007.

[55] Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V., Sterochemistry of ploypeptide chain conformations, *J. Mol. Biol.*, vol. 7, pp. 95–99, 1963.

[56] Ramakrishnan, C. and Ramachandran, G. N., Stereochemical criteria for polypeptide and protein chain conformations II. allowed conformations for a pair of peptide units, *Biophys. J.*, vol. 5, pp. 909–933, 1965.

[57] Kleywegt, G. J. and Jones, T. A., Phi/Psi-chology: Ramachandran revisited, *Structure*, vol. 4, pp. 1395–1400, 1996.

[58] Hovmöller, S., Zhou, T., and Ohlson, T., Conformations of amino acids in proteins, *Acta Cryst.*, vol. D58, pp. 768–776, 2002.

[59] Priestle, J. P., Improved dihedral-angle restraints for protein structure refinement, *J. Appl. Cryst.*, vol. 36, pp. 34–42, 2003.

[60] Sheik, S. S., Ananthalakshmi, P., Bhargavi, G. R., and Sekar, K., CADB: Conformation angles dataBase of proteins, *Nucl. Acids Res.*, vol. 31, pp. 448–451, 2003.

[61] Dayalan, S., Bevinakoppa, S., and Schroder, H., A dihedral angle database of short sub-sequences for protein structure prediction, The Second Asia-Pacific Bioinfomatics Conference, Australian Computer Society, INC., 2004.

[62] Sali, A. and Blundell, T. L., Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.*, vol. 234, pp. 779–815, 1993.

[63] Fiser, A., Gian Do, R. K., and Šali, A., Modeling of loops in protein structures, *Protein Sci.*, vol. 9, pp. 1753–1773, 2000.

[64] Brünger, A. T., Free R value: a novel statistical quantity for assessing the accuracy of crystal structures, *Nature*, vol. 355, pp. 472–475, 1992.

[65] Zheng, Q., Rosenfeld, R., DeLisi, C., and Kyle, D. J., Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations, *Protein Sci.*, vol. 3, pp. 493–506, 1994.

[66] Mathiowetz, A. M. and Goddard III, W. M., Building proteins from C_α coordinates using the dihedral probability grid Monte Carlo method, *Protein Sci.*, vol. 4, pp. 1217–1232, 1995.

[67] Cheng, B., Nayeem, A., and Scheraga, H. A., From secondary structure to three-dimensional structure:improved dihedral angle probability distribution function for use with energy searches for native structures of polypeptides and proteins, *J. Comp. Chem.*, vol. 17, pp. 1453–1480, 1996.

[68] Lovell, S. C., Davis, I. W., Arendall III, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C., Structure validation by C_α geometry: ϕ, ψ and C_β deviation, *Proteins*, vol. 50, pp. 437–450, 2003.

[69] Betancourt, M. R. and Skolnick, J., Local propensities and statistical potentials of backbone dihedral angles in proteins, *J. Mol. Biol.*, vol. 342, pp. 635–649, 2004.

[70] Hooft, R., Vriend, G., Sander, C., and Abola, E., Errors in protein structures, *Nature*, vol. 381, pp. 272–272, 1996.

[71] Vriend, G., WHAT IF: a molecular modelling and drug design program, *J. Mol. Graph.*, vol. 8, pp. 52–56, 1990.

[72] Morris, A. L., MacArthur, M. W., Hutchinson, E. G., and Thornton, J. M., Stereochemical quality of protein structure coordinates, *Proteins*, vol. 12, pp. 345–364, 1992.

[73] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M., PROCHECK: a program to check the stereochemical quality of protein structures, *J. Appl. Cryst.*, vol. 26, pp. 283–291, 1993.

[74] Wu, T. T. and Kabat, E. A., An attempt to locate the non-helical and permissively helical sequences of proteins: application to the variable regions of immunoglobulin light and heavy chains, *Proc. Natl. Acad. Sci. USA*, vol. 68, pp. 1501–1506, 1971.

[75] Kabat, E. A. and Wu, T. T., Construction of a three-dimensional model of the polypeptide backbone of the variable region of kappa immunoglobulin light chains, *Proc. Natl. Acad. Sci. USA*, vol. 69, pp. 960–964, 1972.

[76] Wu, T. T. and Kabat, E. A., Attempt to evaluate influence of neighboring amino-acid (n-1) and (n+1) on backbone conformation of amino-acid (n) in proteins-use in predicting 3-dimensional structure of polypeptide backbone of other proteins, *J. Mol. Biol.*, vol. 75, pp. 13–31, 1973.

[77] Pappu, R. V., Srinivasan, R., and Rose, G. D., The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding, *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 12565–12570, 2000.

[78] Chakrabarti, P. and Pal, D., The interrelationships of side-chain and main-chain conformations in proteins, *Prog. Biophys. Mol. Biol.*, vol. 76, pp. 1–102, 2001.

[79] Zaman, M. H., Shen, M. Y., Berry, R. S., Freed, K. F., and Sosnick, T. R., Investigations into sequence and conformational dependence of backbone entropy,inter-basin dynamics and the Flory isolated-pair hypothesis for peptides, *J. Mol. Biol.*, vol. 331, pp. 693–711, 2003.

[80] Shortle, D., Composite of local structure propensities: Evidence for local encoding of long-rang structure, *Protein Sci.*, vol. 11, pp. 18–26, 2002.

[81] Shortle, D., Propensities, probabilities, and the Boltzmann hypothesis, *Protein Sci.*, vol. 12, pp. 1298–1302, 2003.

[82] Fang, Q. J. and Shortle, D., A Consistent set of statistical potentials for quantifying local side-chain and backbone interactions, *Protein Sci.*, vol. 60, pp. 90–96, 2005.

[83] Fang, Q. J. and Shortle, D., Enhanced Sampling near the Native Conformation Using Statistical Potentials for Local Side-Chain and Backbone Interactions, *Proteins*, vol. 60, pp. 97–102, 2005.

[84] Zhong, S., Moix, J. M., Quirk, S., and Hernandez, R., Dihedral-angle information entropy as a gauge of secondary structure propensity, *Biophys. J.*, vol. 91, pp. 4014–4023, 2006.

[85] Shannon, C. E., A mathematical theory of communication, *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[86] Yang1, W. Y. and Gruebele, M., Folding at the speed limit, *Nature*, vol. 423, pp. 193–197, 2003.

[87] Matouschek, A., Jr, J. K., and Fersht, L. S. A., Mapping the transition state and pathway of protein folding by protein engineering, *Nature*, vol. 340, pp. 122–126, 1989.

[88] Sosnick, T., Dothager, R., and Krantz, B., Differences in the folding transition state of ubiquitin indicated by f and c analyses, *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 17377–17382, 2004.

[89] Schuler, B., Lipman, E., and Eaton, W., Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy, *Nature*, vol. 419, pp. 743–747, 2002.

[90] Magg, C., Kubelka, J., Holtermann, G., Haas, E., and Schmid, F., Specificity of the initial collapse in the folding of the cold shock protein, *J. Mol. Biol.*, vol. 360, pp. 1067–1080, 2006.

[91] Maity, H., Maity, M., Krishna, M., Mayne, L., and Englander, S., Protein folding: the stepwise assembly of foldon units, *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 4741–4746, 2005.

[92] Kim, P. and Baldwin, R., Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding, *Annu. Rev. Biochem.*, vol. 51, pp. 459–489, 1982.

[93] Baldwin, R. L., *Pathways and Mechanism of Protein Folding*, pp. 1–6. New York: Plenum, 1994.

[94] Roder, H., Elove, G., and Englander, S., Structural characterization of folding intermediates in cytochrome c by H-exchange labeling and proton NMR, *Nature*, vol. 335, pp. 700–704, 1988.

[95] Sosnick, T., Mayne, L., Hiller, R., and Englander, S., The barriers in protein folding, *Nat. Struct. Biol.*, vol. 1, pp. 149–156, 1994.

[96] Alber, T., Sun, D., Wilson, K., Wozniak, J., Cook, S., and Matthews, B., Contributions of hydrogen-bonds of Thr-157 to the thermodynamic stability of phage T4 lysozyme, *Nature*, vol. 330, pp. 41–46, 1987.

[97] Shortle, D., Guanidine-hydrochloride denaturation studies of mutant forms of staphylococcal nuclease, *J. Cell. Biochem.*, vol. 30, pp. 281–289, 1986.

[98] Shortle, D. and Ackerman, M., Persistence of native-like topology in a denatured protein in 8 M urea, *Science*, vol. 293, pp. 487–489, 2001.

[99] Jackson, S. and Fersht, A., Folding of chymotrypsin in- hibitor-2.1. Evidence for a two-state transition, *Biochemistry*, vol. 30, pp. 10428–10435, 1991.

[100] Li, L. and Shakhnovich, E., Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations, *Proc. Natl. Acad. Sci. USA*, vol. 8, pp. 13014–13018, 2001.

[101] Bryngelson, J. and Wolynes, P., Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. USA*, vol. 84, pp. 7524–7528, 1987.

[102] Bryngelson, J. and Wolynes, P., Intermediates and bar- rier crossingin a random energy model (with applications to protein folding), *J. Phys. Chem.*, vol. 93, pp. 6902–6915, 1989.

[103] Bryngelson, J. and Wolynes, P., A simple statistical field theory of heteropolymer collapse with application to protein folding, *Biopolymers*, vol. 30, pp. 177–188, 1990.

[104] Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., and Socci, N. D., Toward an outline of the topography of a realistic protein folding funnel, *Proc. Natl. Acad. Sci. USA*, vol. 92, pp. 3626–3630, 1995.

[105] Bryngelson, J., Onuchic, J., Socci, N., and Wolynes, P., Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis, *Proteins: Struct. Func. Gen.*, vol. 21, pp. 167–195, 1995.

[106] Sali, A., Shakhnovich, E., and Karplus, M., Kinetics of protein folding: A lattice model study of the requirements for folding to the native state, *J. Mol. Biol.*, vol. 235, pp. 1614–1636, 1994.

[107] Chan, H. S. and Dill, K. A., Polymer principles in protein structure and stability, *Annu. Rev. Biophys. Biomol. Struct.*, vol. 20, pp. 447–490, 1991.

[108] Socci, N. D. and Onuchic, J. N., Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique, *J. Chem. Phys.*, vol. 103, pp. 4732–4744, 1995.

[109] Wolynes, P. G., Onuchic, J. N., and Thirumalai, D., Navigating the folding routes, *Science*, vol. 267, pp. 1619–20, 1995.

[110] Socci, N. D., Onuchic, J. N., and Wolynes, P. G., Diffusive dynamics of the reaction coordinate for protein folding funnels, *J. Chem. Phys.*, vol. 104, pp. 5860–5868, 1996.

[111] Dill, H. C. . K. A., Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics, *Proteins*, vol. 30, pp. 2–33, 1998.

[112] Dinner, A. R. and Karplus, M., The thermodynamics and kinetics of protein folding: A lattice model analysis of multiple pathways with intermediates, *J. Phys. Chem. B*, vol. 103, pp. 7976–7994, 1999.

[113] Locker, C. R. and Hernandez, R., Folding behavior of model proteins with weak energetic frustration, *J. Chem. Phys.*, vol. 120, pp. 11292–11303, 2004.

[114] Das, P., Matysiak, S., and Clementi, C., Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes, *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 10141–10146, 2005.

[115] Das, P., Moll, M., Stamati, H., Kavraki, L., and Clementi, C., Low-dimensional, free-energy landscapes of protein folding reactions by nonlinear dimensionality reduction, *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 9885–9890, 2006.

[116] Zhou, R., Berne, B., and Germain, R., The Free Energy Landscape for Beta Hairpin Folding in Explicit Water, *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 14931–14936, 2001.

[117] Mayor, U., Johnson, C. M., Daggett, V., and Fersht, A. R., Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation, *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 13518–13522, 2000.

[118] Jong, D. D., Riley, R., Alonso, D. O., and Dagget, V., Probing the Energy Landscape of Protein Folding/Unfolding Transition States, *J. Mol. Biol.*, vol. 319, pp. 229–242, 2002.

[119] Zwanzig, R., Szabo, A., and Bagchi, B., Levinthal's paradox, *Proc. Natl. Acad. Sci. USA*, vol. 89, pp. 20–22, 1992.

[120] Karplus, M., The Levinthal paradox: yesterday and today, *Fold. Des.*, vol. 2, pp. 69–75, 1997.

[121] Alder, B. and Wainright, T., Phase Transition for a Hard Sphere System, *J. Chem. Phys.*, vol. 27, p. 1208, 1957.

[122] Rahman, A., Correlations in the Motion of Atoms in Liquid Argon, *Phys. Rev. A*, vol. 136, p. 405, 1964.

[123] Rahman, A. and Stillinger, F., Molecular Dynamics Study of Liquid Water, *J. Chem. Phys.*, vol. 55, p. 3336, 1971.

[124] McCammon, J., Gelin, B., and Karplus, M., Dynamics of Folded Proteins, *Nature*, vol. 267, p. 585, 1977.

[125] Itzhaki, L., Otzen, D., and Fersht, A., The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding, *J. Mol. Biol.*, vol. 254, pp. 260–288, 1995.

[126] Daura, X., Jaun, B., Seebach, D., van Gunsteren, W., and Mark, A., Reversible peptide folding in solution by molecular dynamics simulation, *J. Mol. Biol.*, vol. 280, pp. 925–932, 1998.

[127] Day, R. and Daggett, V., Increasing Temperature Accelerates Protein Unfolding Without Changing the Pathway of Unfolding, *J. Mol. Biol.*, vol. 322, pp. 189–203, 2002.

[128] Sugita, Y. and Okamoto, Y., Replica-exchange molecular dynamics method for protein folding, *Chem. Phys. Lett.*, vol. 314, pp. 141–151, 1999.

[129] Sugita, Y., Kitao, A., and Okamoto, Y., Multidimensional replica-exchange method for free-energy calculations, *J. Chem. Phys.*, vol. 113, p. 6042, 2000.

[130] Okur, A., Roe, D. R., Cui, G., Hornak, V., and Simmerling, C., Improving Convergence of Replica-Exchange Simulations through Coupling to a High-Temperature Structure Reservoir, *J. Chem. Theory Comput.*, vol. 3, pp. 557–568, 2007.

[131] Darve, E. and Pohorille, A., Calculating free energies using average force, *J. Chem. Phys.*, vol. 115, pp. 9169–9183, 2001.

[132] Darve, E., Rodriguez-Gomez, D., and Pohorille, A., Adaptive biasing force method for scalar and vector free energy calculations, *J. Chem. Phys.*, vol. 128, p. 144120, 2008.

[133] Chandler, D., *Introduction to modern statistical mechanics*. New York: Oxford, 1987.

[134] Jarzynski, C., Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach, *Phys. Rev. E*, vol. 56, pp. 5018–5035, 1997.

[135] Jarzynski, C., Nonequilibrium equality for free energy differences, *Phys. Rev. Lett.*, vol. 78, pp. 2690–2693, 1997.

[136] Park, S. and Schulten, K., Calculating potentials of mean force from steered molecular dynamics simulations, *J. Chem. Phys.*, vol. 120, pp. 5946 – 5961, 2004.

[137] Xiong, H., Crespo, A., Marti, M., Estrin, D., and Roitberg, A. E., Free energy calculations with non-equilibrium methods: applications of the Jarzynksi relationship, *Theor. Chem. Acta*, vol. 116, pp. 338–346, 2006.

[138] Torras, J., de M. Seabra, G., and Roitberg, A. E., A Multiscale Treatment of Angeliś Salt Decomposition, *J. Chem. Theory Comput.*, vol. 5, pp. 37–46, 2009.

[139] Piccinini, E., Ceccarelli, M., Affinito, F., Brunetti, R., and Jacoboni, C., Biased Molecular Simulations for Free-Energy Mapping: A Comparison on the KcsA Channel as a Test Case, *J. Chem. Theory Comput.*, vol. 4, pp. 173–183, 2008.

[140] Huang, H., Ozkirimli, E., and Post, C. B., Comparison of Three Perturbation Molecular Dynamics Methods for Modeling Conformational Transitions, *J. Chem. Theory Comput.*, vol. 5, pp. 1304–1314, 2009.

[141] Liphardt, J., Dumont, S., Smith, S. B., Jr., I. T., and Bustamante, C., Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality, *Science*, vol. 296, pp. 1832–1835, 2002.

[142] Douarche, F., Ciliberto, S., Petrosyan, A., and Rabbiosi, L., An experimental test of the Jarzynski equality in a mechanical experiment, *Europhys. Lett.*, vol. 70, pp. 593–599, 2005.

[143] Ozer, G., Valeev, E., Quirk, S., and Hernandez, R., Adaptive steered molecular dynamics of the long-distance unfolding of Neuropeptide Y, *J. Chem. Theory Comput.*, vol. 6, pp. 3026–3038, 2010.

[144] Sudarsanam, S., DuBose, R. F., March, C. J., and Srinivasan, S., Modeling protein loops using a ϕ_{i+1},ψ_i dimer database, *Protein Sci.*, vol. 4, pp. 1412–1420, 1995.

[145] Sudarsanam, S. and Srinivasan, S., Searching for protein loops in parallel, *CABIOS*, vol. 11, pp. 591–593, 1995.

[146] Sudarsanam, S. and Srinivasan, S., Sequence-dependent conformational sampling using a database of ϕ_{i+1} and ψ_i angles for predicting polypeptide backbone conformations, *Protein Eng.*, vol. 10, pp. 1155–1162, 1997.

[147] Parker, J. M. R., The relationship between peptide plane rotation (PPR) and similar conformations, *J. Comp. Chem.*, vol. 20, pp. 947–955, 1999.

[148] Hodel, A., Kautz, R. A., Jacobs, M. D., and Fox, R. O., Stress and strain in staphylococcal nuclease, *Protein Sci.*, vol. 2, pp. 838–850, 1993.

[149] Hodel, A., Kautz, R. A., and Fox, R. O., Stabilization of a strained protein loop conformation through protein engineering, *Protein Sci.*, vol. 4, pp. 484–495, 1995.

[150] Crooks, G. E., Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems, *J. Stat. Phys.*, vol. 90, pp. 1481–1487, 1998.

[151] Hummer, G. and Szabo, A., Free energy reconstruction from nonequilibrium single-molecule pulling experiments, *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 3658–3661, 2001.

[152] Zimanyi, E. N. and Silbey, R. J., The work-Hamiltonian connection and the usefulness of the Jarzynski equality for free energy calculations, *J. Chem. Phys.*, vol. 130, p. 171102, Jan 2009.

[153] Park, S., Khalili-Araghi, F., Tajkhorshid, E., and Schulten, K., Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality, *J. Chem. Phys.*, vol. 119, pp. 3559 – 3566, 2003.

[154] Amaro, R., Tajkhorshid, E., and Luthey-Schulten, Z., Developing an energy landscape for the novel function of a (beta/alpha)8 barrel: Ammonia conduction through HisF, *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 7599 – 7604, 2003.

[155] Echeverria, I. and Amzel, L. M., Helix propensities calculations for amino acids in alanine based peptides using Jarzynskis equality, *Proteins: Struct. Func. Bioinfo.*, vol. 78, pp. 1302–1310, 2010.

[156] Tucker, A. K. and Hernandez, R., Observation of a trapping transition in the diffusion of a thick needle through fixed point scatterers, *J. Phys. Chem. B*, vol. 114, pp. 9628–9634, 2010.

[157] Zwanzig, R. W., High temperature equation of state by a perturbation method. i. nonpolar gases, *J. Chem. Phys.*, vol. 22, pp. 1420–1426, 1954.

[158] Straatsma, T. and McCammon, A., Computational alchemy, *Annu. Rev. Phys. Chem.*, vol. 43, pp. 407–435, 1992.

[159] Rodriguez-Gomez, D., Darve, E., and Pohorille, A., Assessing the efficiency of free energy calculation methods, *J. Chem. Phys.*, vol. 120, pp. 3563–3578, 2004.

[160] Geissler, P. L. and Dellago, C., Equilibrium time correlation functions from irreversible transformations in trajectory space, *J. Phys. Chem. B*, vol. 108, pp. 6667–6672, 2004.

[161] Ytreberg, F. M. and Zuckerman, D. M., Single-ensemble nonequilibrium path-sampling estimates of free energy differences, *J. Chem. Phys.*, vol. 120, pp. 10876–10879, 2004.

[162] Hatano, T. and Sasa, S. I., Steady-state thermodynamics of Langevin systems, *Phys. Rev. Lett.*, vol. 86, pp. 3463–3466, 2001.

[163] Wu, D. and Kofke, D. A., Model for small-sample bias of free-energy calculations applied to Gaussian-distributed nonequilibrium work measurements, *J. Chem. Phys.*, vol. 121, pp. 8742–8747, 2004.

[164] Eaton, W. A., Muoz, V., Thompson, P. A., Henry, E. R., and Hofrichter, J., Kinetics and Dynamics of Loops, alpha-Helices, beta-Hairpins, and Fast-Folding Proteins, *Acc. Chem. Res.*, vol. 31, pp. 745–753, 1998.

[165] Pande, V. S. and Rokhsar, D. S., Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G, *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 9062–9067, 1999.

[166] Dinner, A. R., Lazaridis, T., and Karplus, M., Understanding beta-hairpin formation, *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 9068–9073, 1999.

[167] Crane, J. C., Koepf, E. K., Kelly, J. W., and Gruebele, M., Mapping the transition state of the WW domain beta-sheet, *J. Mol. Biol.*, vol. 298, pp. 283–292, 2000.

[168] Beke-Somfai, T. and Perczel, A., Zipper-Like Unfolding of beta-Sheets Accessed by Pioneer Water Molecules: Atomic Resolution of Forced Unfold Reveals Different Mechanisms for Parallel and Antiparallel Motifs, *J. Phys. Chem. Lett.*, vol. 1, pp. 1341–1345, 2010.

[169] Spichty, M., Cecchini, M., and Karplus, M., Conformational Free-Energy Difference of a Miniprotein from Nonequilibrium Simulations, *J. Phys. Chem. Lett.*, vol. 1, pp. 1922–1926, 2010.

[170] Minh, D. D. L. and McCammon, J. A., Springs and Speeds in Free Energy Reconstruction from Irreversible Single-Molecule Pulling Experiments, *J. Phys. Chem. B*, vol. 112, pp. 5892 – 5897, 2008.

[171] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K., Scalable molecular dynamics with NAMD, *J. Comp. Chem.*, vol. 28, pp. 1781–1802, 2005.

[172] Brooks, B., Bruccoleri, R., Olafson, R., States, D., Swaminathan, S., and Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *J. Comp. Chem.*, vol. 4, pp. 187–217, 1983.

[173] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L., Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.*, vol. 79, pp. 926–935, 1983.

[174] Gray, T. and Morley, J., Neuropeptide Y: Anatomical distribution and possible function in mammalian nervous system, *Life Sci.*, vol. 38, pp. 389–401, 1986.

[175] Dumont, Y., Fournier, A., and Quirion, R., Neuropeptide Y and neuropeptide Y receptor subtypes in brain and peripheral tissues, *Progr. Neurobiol.*, vol. 38, pp. 125–167, 1992.

[176] Turton, M., O'Shea, D., and Bloom, S., *Central effects of neuropeptide Y with emphasis on its role in obesity and diabetes*, pp. 15–39. San Diego, CA: Academic Press, 1997.

[177] Larhammar, D., Structural diversity of receptors for neuropeptide Y, peptide YY and pancreatic polypeptide, *Regul. Pept.*, vol. 65, pp. 165–174, 1996.

[178] Wraith, A., Tornsten, A., Chardon, P., Harbitz, I., Chowdhary, B. P., Andersson, L., Lundin, L.-G., and Larhammar, D., Evolution of the neuropeptide Y receptor family: Gene and Chromosome duplications deduced from the cloning and mapping of the five receptor subtype genes in pig, *Genome Res.*, vol. 10, pp. 302–310, 2000.

[179] Larhammar, D., Evolution of neuropeptide Y, peptide YY, and pancreatic polypeptide, *Regul. Pept.*, vol. 62, pp. 1–11, 1996.

[180] Li, X., Sutcliffe, M. J., Schwartz, T. W., and Dobson, C. M., Sequence-specific proton NMR assignments and solution structure of bovine pancreatic polypeptide, *Biochemistry*, vol. 31, pp. 1245–1253, 1992.

[181] Darbon, H., Bernassau, J., Deleuze, C., Chenu, J., Roussel, A., and Cambillau, C., Solution conformation of human neuropeptide Y by 1H nuclear magnetic resonance and restrained molecular dynamics, *Eur. J. Biochem.*, vol. 209, pp. 765–771, 1992.

[182] Nordmann, A., Blommers, M., Fretz, H., Arvinte, T., and Drake, F., Aspects of the molecular structure and dynamics of neuropeptide Y, *Eur. J. Biochem.*, vol. 261, pp. 216–226, 1999.

[183] Cowley, D., Hoflack, J., Pelton, J., and Saudek, V., Structure of neuropeptide Y dimer in solution, *Eur. J. Biochem.*, vol. 205, pp. 1099–1106, 1992.

[184] Mierke, D., Durr, H., Kessler, H., and Jung, G., Neuropeptide Y: Optimized solid-phase synthesis and conformational analysis in trifluoroethanol, *Eur. J. Biochem.*, vol. 206, pp. 39–48, 1992.

[185] Monks, S., Karagianis, G., Howlett, G., and Norton, G., Solution structure of human neuropeptide Y, *J.Biomol.NMR*, vol. 8, pp. 379–390, 1996.

[186] Bader, R., Bettio, A., Beck-Sickinger, A. G., and Zerbe, O., Structure and dynamics of micelle-bound neuropeptide Y: Comparison with unligated NPY and implications for receptor selection, *Genome Res.*, vol. 10, pp. 302–310, 2000.

[187] Lerch, M., Mayrhofer, M., and Zerbe, O., Structural similarities of micelle-bound peptide YY (PYY) and neuropeptide Y (NPY) are related to their affinity profiles at the Y receptors, *J. Mol. Biol.*, vol. 339, pp. 1153–1168, 2004.

[188] Bettio, A., Dinger, M. C., and Beck-Sickinger, A. G., The neuropeptide Y monomer in solution is not folded in the pancreatic-polypeptide fold, *Protein Sci.*, vol. 11, pp. 1834–1844, 2002.

[189] Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P., and Wu, C.-W., X-ray analysis (1. 4-A resolution) of avian pancreatic polypeptide: Small globular protein hormone, *Proc. Natl. Acad. Sci. USA*, vol. 78, pp. 4175–4179, 1981.

[190] Daggett, V. and Fersht, A. R., Is there a unifying mechanism for protein folding?, *Trends Biochem. Sci.*, vol. 28, pp. 18–25, 2003.

[191] Park, P. J. and Lee, S., Particle mesh Ewaldi: An N log(N) method for Ewald sums in large systems, *J. Chem. Phys.*, vol. 98, pp. 10089–10092, 1993.

[192] Wang, T. and Wade, R. C., On the Use of Elevated Temperature in Simulations To Study Protein Unfolding Mechanisms, *J. Chem. Theory Comput.*, vol. 3, pp. 1476–1483, 2007.

[193] Sylte, I., Andrianjara, C., Calvet, A., Pascal, Y., and Dahl, S., Molecular dynamics of NPY Y1 receptor activation, *Bioorg. Med. Chem.*, vol. 7(12), pp. 2737–2748, 1999.

[194] Ertekin, A., Nussinov, R., and Haliloglu, T., Association of putative concave protein-binding sites with the fluctuation behavior of residues, *Protein Sci.*, vol. 15, p. 22652277, 2006.

[195] Hänggi, P., Talkner, P., and Borkovec, M., Reaction-rate theory: Fifty years after Kramers, *Rev. Mod. Phys.*, vol. 62, pp. 251–341, 1990. and references therein.

[196] Pollak, E. and Talkner, P., Reaction rate theory: What it was, where it is today, and where is it going?, *Chaos*, vol. 15, pp. 026116–1–11, 2005.

[197] Hernandez, R., Bartsch, T., and Uzer, T., Transition state theory in liquids beyond planar dividing surfaces, *Chem. Phys.*, vol. 370, pp. 270–276, 2010.

[198] Beck-Sickinger, A. G., Wieland, H. A., Wittneben, H., Willim, K.-D., Rudolf, K., and Jung, G., Complete L-alanine scan of neuropeptide Y reveals ligands binding to Y1 and Y2 receptors with distinguished conformations, *Eur. J. Biochem.*, vol. 225(3), pp. 947–958, 1994.

[199] Fournier, A., Gagnon, D., Quirion, R., Dumont, Y., Pheng, L.-H., and St-Pierre, S., Conformational and biological studies of neuropeptide Y analogs containing structural alterations, *Mol. Pharmacol.*, vol. 45, pp. 93–101, 1994.

[200] Best, R. B., Li, B., Steward, A., Daggett, V., and Clarke, J., Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation, *Biophys. J.*, vol. 81, pp. 2344–2356, 2001.

[201] Marcinkiewicz, J., Sur une propriete de la loi de Gauss, *Math. Z*, vol. 44, pp. 612 – 618, 1939.

[202] Kubelka, J., Hofrichter, J., and Eaton, W. A., The protein folding speed limit, *Curr. Opin. Struct. Biol.*, vol. 14, pp. 76–88, 2004.

[203] Neumoin, A., Mares, J., Lerch-Bader, M., Bader, R., and Zerbe, O., Probing the formation of stable tertiary structure in a model miniprotein at atomic resolution: Determinants of stability of a helical hairpin, *J. Am. Chem. Soc.*, vol. 129, pp. 8811–8817, 2007.

[204] Keire, D., Kobayashi, M., Solomon, T., and Jr., J. R., Solution structure of monomeric peptide YY supports the functional significance of the PP-fold, *Biochemistry*, vol. 39, pp. 9935–9942, 2000.

[205] Gellman, M. W. S., Backbone thioester exchange: A new approach to evaluating higher order structural stability in polypeptides, *J. Am. Chem. Soc.*, vol. 126, pp. 11172–11174, 2004.

[206] Neumoin, A., Mares, J., Lerch-Bader, M., Bader, R., and Zerbe, O., Probing the formation of stable tertiary structure in a model miniprotein at atomic resolution: Determinants of stability of a helical hairpin, *J. Am. Chem. Soc.*, vol. 129, pp. 8811–8817, 2007.

[207] Hodges, A. and Schepartz, A., Engineering a monomeric miniature protein, *J. Am. Chem. Soc.*, vol. 129, pp. 11024–11025, 2007.

[208] Amunson, K., Ackels, L., and Kubelka, J., Site-specific unfolding thermodynamics of a helix-turn-helix protein, *J. Am. Chem. Soc.*, vol. 130, pp. 8146–8147, 2008.

[209] Du, D. and Gai, F., Understanding the folding mechanism of an R-helical hairpin, *Biochemistry*, vol. 45, pp. 13131–13139, 2006.

[210] Waegele, M. M. and Gai, F., Infrared Study of the Folding Mechanism of a Helical Hairpin: Porcine PYY, *Biochemistry*, vol. 49, pp. 7659–7664, 2010.

[211] Dyer, R., Gai, F., Woodruff, W., Gilmanshin, R., and Callender, R., Infrared studies of fast events in protein folding, *Acc. Chem. Res.*, vol. 31, pp. 709–716, 1998.

[212] Fersht, A., Matouschek, A., and Serrano, L., The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding, *J. Mol. Biol.*, vol. 224, pp. 771–782, 1992.

[213] Heus, H., van Vugt-Jonker, A., Zimmermann, J., Puchner, E., and Gaub, H., AFM-based single-molecule force spectroscopy of RNA unfolding, *Anal. Biochem.*, vol. yyy, pp. xxx–xxx, 2011.

[214] Onoa, B., Dumont, S., Liphardt, J., Smith, S. B., Jr., I. T., and Bustamante, C., Identifying Kinetic Barriers to Mechanical Unfolding of the T. thermophila Ribozyme, *Science*, vol. 299, pp. 1892–1895, 2003.

[215] Cecconi, C., Shank, E. A., Bustamante, C., and Marqusee, S., Direct Observation of the Three-State Folding of a Single Protein Molecule, *Science*, vol. 309, pp. 2057–2060, 2005.

[216] Cecconi, C., Shank, E., Dahlquist, F., Marqusee, S., and Bustamante, C., Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers, *Eur. Biophys. J.*, vol. 37, pp. 729–738, 2008.

[217] Dryden, S., Pickavance, L., Frankish, H. M., and Williams, G., Increased neuropeptide Y secretion in the hypothalamic paraventricular nucleus of obese (fa/fa) Zucker rats, *Brain Res.*, vol. 690, pp. 185–188, 1996.

[218] Kitlinska, J., Kuo, L., Abe, K., Yu, J. P. M., Li, L., Tilan, J., Toretsky, J., and Zukowska, Z., Role of neuropeptide Y and dipeptidyl peptidase IV in regulation of Ewing's sarcoma growth, *Advr Exp. Med. Biol.*, vol. 690, pp. 185–188, 1996.

[219] Kitlinska, J., Kuo, L., Abe, K., Yu, J. P. M., Li, L., Everhart, J. T. L., Lee, E., Zukowska, Z., and Toretsky, J., Differential effects of neuropeptide Y on the growth and vascularization of neural crest-derived tumors, *Cancer Res.*, vol. 65, pp. 1719–1728, 2005.